Predicting Cognitive Presence in At-Scale Online Learning: MOOC and For-Credit Online Course Environments

Jeonghyun Lee
Farahnaz Soleimani
India Irish
John Hosmer IV
Meryem Yilmaz Soylu
Roy Finkelberg
Saurabh Chatterjee
Georgia Institute of Technology, USA

Abstract
In this study, we work towards a strategy to measure and enhance the quality of interactions in discussion forums at scale. We present a machine learning (ML) model which identifies the phase of cognitive presence exhibited by a student's post and suggest future applications of such a model to help online students develop higher-order thinking. We collect discussion forum transcript data from two online courses: CS1301 (an introductory computer programming MOOC) offered by edX and CS6601 (a graduate course on artificial intelligence) which uses the Piazza online discussion tool. We manually code a random sample of students' posts based on the Community of Inquiry coding scheme and explore trends in cognitive presence within and across the courses. We further use this coded data to analyze the relationship between students' observed cognitive presence and course grades. In terms of testing and building an ML model, we use a Bidirectional Encoder Representations from Transformers model that uses a deep learning technique to train large text corpus and fine-tune the language model. Our results suggest that deeper cognitive engagement with course concepts, as expressed by higher cognitive presence, are associated with better learning outcomes for students in both course settings. Our ML approach achieves 92.5% accuracy on the classification task, motivating the use of ML for instructional interventions in online courses. We expect that our research study will not only contribute to extending the literature on cognitive presence but also have a beneficial impact on online instructors or curriculum developers in higher education.

Keywords: Cognitive presence, discussion forums, machine learning, higher education

Lee, J., Soleimani., F., Irish, I., Hosmer, J. Soylu, M., Y., Finkelberg, R., & Chatterjee, S. (2022). Predicting cognitive presence in at-scale online learning: MOOC and for-credit online course environments, *Online Learning, 26*(1), 58-79. DOI: 10.24059/olj.v26i1.3060

In this study, we explore how students develop higher-order thinking through participation in online discussion forums by adopting the community of inquiry (CoI) framework (Garrison, Anderson, & Archer, 2001). This conceptual framework has been widely used to guide research in educational experiences of students situated in various collaborative online learning environments such as asynchronous discussion forums (Galikyan, Admiraal, & Kester, 2021; Garrison, Anderson, & Archer, 2010). Specifically, we compare trends of students' cognitive presence between two different online course settings: an undergraduate-level massive open online course (MOOC) that is accessible to the public free of charge, and a graduate-level course which is part of an online degree program. Further, we explore the idea of automatically identifying students' levels of critical thinking from discussion forum transcripts. We present the application of a machine learning classification model for natural language processing which identifies the phase of cognitive presence observed in a student's forum post and suggest future applications of such systems to CoI-based interventions.

Review of Related Literature

Learning in Asynchronous Discussion Forums

Asynchronous discussion forums serve as a platform to support the learning process of online students by allowing them to build and share knowledge with others. Regarding learning in MOOCs, several studies have revealed that instructors perceive the beneficial role of online discussion features in facilitating quality learning (e.g., Askeroth & Richardson, 2019). Asynchronous discussion platforms are usually designed to help students learn from others by not only providing a venue for communication and interaction among students and instructors but also by enhancing content delivery (Baglione & Nastanski, 2007). Previous research suggests that various factors associated with the affordances of asynchronous discussion forums can impact students' participation in online discussion. Such factors include relational capital among participants (Chapman, Storberg-Walker, & Stone, 2008), visibility of social cues (Cheung, Hew, & Ng, 2008), and instructors' presence (An, Shin, & Lim, 2009; Baran & Correia, 2009). Other factors can also mediate between students' engagement with online discussion and learning. For example, participation in online discussions at a deep level (e.g., reflecting, refining meaning) has been found to be related to high academic achievement (Bliuc et al., 2009; Galikyan et al., 2021). Thus, it is critical to design online discussion environments that sustain a sense of community and support students socially and cognitively.

Despite its beneficial influence on students' learning, online discussion forums pose some challenges in terms of promoting active participation among students, effectively facilitating conversation, organizing an optimal structure for co-constructing knowledge, and dealing with time constraints commonly confronted by instructors (Mazzolini & Maddison, 2007; Zhu, Bonk, & Sari, 2018). To overcome these challenges, deNoyelles, Zydney and Chen (2014) proposed a list of strategies for instructors based on the CoI framework. For example, an instructor can use social modeling cues (e.g., calling a student by name), graded discussion assignments, discussion prompts, facilitation techniques (e.g., questioning), modest feedback (e.g., posting less often but in a meaningful way) and protocol prompts with structured goals and roles in a specific deadline. Beyond these strategies, the purposeful design of online platform interfaces (Quintana, Pinto, & Tan, 2021; Zhu et al., 2018) and implementation of instructional strategies to improve students' cognitive engagement (Garrison & Akyol, 2015; Kilis & Yildirim, 2019) have been shown to enhance successful and engaging online learning.

Cognitive Presence in Online Learning Contexts

According to the CoI framework, collaborative knowledge construction can be fostered through the critical dimensions of teaching, social, and cognitive presence. Previous research has stressed the importance of facilitating cognitive presence to help students engage with critical thinking and deepen their inquiry process in online courses (Garrison, Anderson, & Archer, 2010; Sadaf & Olesova, 2017; Shea & Bidjerano, 2009). From the perspective of the practical inquiry model (i.e., the model of critical thinking), which serves as the theoretical basis of CoI, our study focuses on measuring online students' levels of cognitive presence which can be manifested in four phases, including: triggering event (phase 1), exploration of ideas (phase 2), integration of the ideas generated in the exploratory phase (phase 3), and resolution of the problem or issue (phase 4) (Garrison et al., 2001; Sadaf & Olesova, 2017). Among these four phases of cognitive presence, the phase of integration has been found as the most difficult to detect because it is often difficult to catalyze the advancement from the exploration phase without appropriate support from instructors or advanced peers (Garrison et al., 2001).

A substantial body of research has provided helpful insights into facilitating high levels of cognitive presence in online learning contexts. Some researchers have emphasized the beneficial impact of case-based discussions in which students engage with real-life cases and authentic problem-solving processes (Guo et al. 2021; Kilis & Yildirim, 2019; Sadaf, Kim, & Wang, 2021; Sadaf & Olesova, 2017). Other researchers have stressed the importance of designing online course features that can create an "optimal social space" in which learners share their resources and experiences and develop supportive social networks (Amemado & Manca, 2017). Similarly, Darabi et al. (2011) found that, compared to the traditional approach of asking unstructured probing questions, strategies of scaffolding in which student mentors raise questions that focus on advancing the discussion towards a consensus for finding a solution appeared to help facilitate cognitive presence.

Yet, despite the valuable knowledge gained regarding the facilitation of deeper cognitive engagement of online students, further research is required to understand the impact of cognitive presence on actual learning outcomes (Sadaf et al., 2021). Moreover, extant research has heavily focused on small-scale and for-credit online courses. In fact, researchers have identified challenges of promoting in-depth online discussions, especially in low-stakes MOOC environments with high student drop-out rates (Gao, Zhang, & Franklin, 2013; Hew & Cheung, 2014; Nandi, Hamilton, & Harland, 2012). Moreover, instructors in large-scale online courses are likely to feel overwhelmed by students' posts and struggle to measure the quality of their interactions. Taking these factors into account, we aim to explore the development of cognitive presence observed in discussion forum posts in two different types of online courses—an undergraduate-level MOOC and for-credit online master's course—and examine the relationship between cognitive presence and learning outcomes within each course setting.

Application of Machine Learning and Learning Analytics to Educational Data

Although we can draw meaningful implications about online students' cognitive engagement from the CoI framework, challenges remain with respect to common practices for implementing the CoI coding scheme due to its subjective and manual nature. For instance, the conventional coding process to identify the four phases of cognitive presence typically requires systematic training and time commitment from coders to ensure the reliability of text data interpretation. This can be problematic, especially when analyzing large-scale forum data because of time and resource constraints. To address this problem, we explored machine learning

algorithms and their related natural language processing techniques to create a scalable language model that can train the coding scheme and ultimately predict the cognitive presence of a large amount of discussion forum posts within a short period of time.

Online educational platforms are well suited to apply machine learning techniques because of the massive amount of data being collected for learning (see Appendix). Previous research has used data in the field of education to test the performance and accuracy of various machine learning models designed to discover hidden and complex patterns in online students' learning behaviors (Al-Shabandar et al., 2019; Hew et al., 2020). These efforts have encouraged the community to continue utilizing technical but interdisciplinary approaches to better support educational environments. Closely related to these research efforts, the notion of learning analytics has become increasingly popular in higher education settings. Learning analytics has been generally defined as the measurement, collection, analysis and reporting of data about learners and learning environments for purposes of understanding and optimizing the learning process (Siemens & Long, 2011). With higher education institutions being a part of the digital age by integrating online platforms in their learning environment, large data sets are now available throughout the learning process. Researchers have used various learning analytics techniques such as classification, clustering, and text mining (Leitner, Khalil, & Ebner, 2017). These techniques have been used to detect student behavior and predict student performance (Al-Shabandar et al., 2019), identify students at risk (Chen et al., 2018), analyze students' forum interactions, and provide visualization to inform instructors and other key stakeholders (Authors, 2020). However, more research is needed to understand how learning analytics helps improve online instructional practices and students' learning outcomes (Viberg et al., 2018). This encourages researchers to explore other measures, such as cognitive presence, to predict students' performance in online learning environments.

Our work is motivated by the recent trend of applying educational theoretical frameworks and machine learning to understand students' cognitive presence in discussion forums. For example, several studies explored a set of linguistic features of online discussion messages (e.g., LIWC, Coh-Metrics, word embedding similarity) to test which features have predictive relationships with cognitive presence; based on this information researchers developed machine learning models that can automatically classify the level of cognitive presence in the data (e.g., Kovanović et al., 2016; Neto et al., 2021). Similarly, in another study (Hayati, Idrissi, & Bennani, 2020), the authors used text mining and machine learning algorithms to classify students into one of four levels of cognitive engagement including passive, active, constructive, and interactive (Chi & Wylie, 2014) based on their level of cognitive presence and social interactions within discussion forums. Our work intends to offer a technique that examines quality interaction measured by a critical thinking framework to better understand students' learning outcomes. We also acknowledge scalability by designing a model that can exist in low- and high-stakes education environments at scale (Pelánek, 2020).

Research Questions

Our study was guided by three research questions. First, how do online students develop cognitive presence in two different course settings? Second, to what extent does cognitive presence contribute to enhancing students' course grades? Third, can we develop a ML model to detect the level of cognitive presence in discussion forum posts?

Method

Participants and Settings

Our study focused on analyzing discussion forum data collected from two online courses, including an introductory undergraduate-level computer programming MOOC (CS1301) and an online master's degree course about artificial intelligence (CS6601). The CS1301 MOOC is available free of charge to anyone who has signed up for the edX platform. According to the course description, knowledge of basic arithmetic and high school-level algebra is desirable; however, no prior knowledge of computer programming is required from students. Thousands of students are typically enrolled in this low-stakes course; for example, we observed nearly 45,000 students who were enrolled during the Fall 2017 semester. On the other hand, CS6601 has a much smaller class size than CS1301 (e.g., 796 students in Spring 2020), and it is considered a high-stakes for-credit course. The course requires prior knowledge of college-level mathematical concepts and computer programming and algorithms. As one of the core courses in the Online Master's in Computer Science program, CS6601 is designed to incorporate intensive readings, assignments, and independent work. These two courses were taught by various instructors and offered by the same institution—a technology-focused public university in the US.

Data Sources and Procedures

Regarding data collection from CS1301, the research team was provided with the securely encrypted course data from edX, which consisted of course enrollment and participation information from users who have accepted the terms of edX's Privacy Policy. The data were also compliant with the General Data Protection Regulation (GDPR) law, which protects the privacy rights of E.U. residents. Prior to any data analysis, all identifying information was removed from raw data, including usernames within the discussion forum transcript data. For CS6601, we proceeded with data collection based on the institutional review board (IRB)-approved study protocol. We obtained informed consent from the instructor of CS6601 who agreed to provide fully anonymized Piazza transcript data for the purpose of research. Student demographic information was not collected because it was beyond the scope of our present research.

Data sources consisted of a total of 2,341 posts that came from two sets of anonymized transcript data collected from each of the two courses (see Table 1). The CS1301 data was collected during the Fall 2017 and Fall 2018 semesters via edX, a major MOOC provider. This includes a total of 848 comments that were pulled through a stratified random sampling technique. The stratification was based on the number of total comments posted within a certain discussion thread in order to capture the dynamic nature of conversation flows across the discussion board. Regarding CS6601, which was taught during the Spring 2020 semester, we analyzed 1,493 posts collected through the Piazza discussion forum tool. The CS6601 dataset consisted of randomly sampled posts associated with two specific assignments which elicited the most active participation in online discussion. In both courses, participation in the discussion forum was voluntary and was not counted for final grading. However, students in CS6601 were encouraged to post questions to Piazza prior to scheduling an office hours appointment.

Table 1
Description of Student Participation in Online Discussion Forums

	CS1301 (MOOC)	**CS6601 (For-Credit Course)**
Total Number (#) of Posts Coded	848	1,493
Total # of Student Contributors	362	186
Total # of Instructor (TA) Contributors	1 (1)	1 (13)
Total # of Discussion Threads Generated	350	155
Average # of Posts per Thread	2.4	9.6
Average # of Posts per Student	2.5	5.7

Measures

The key measures used in this study include indicators of cognitive presence and final course grades (numeric scores). To measure the cognitive presence of students in discussion forums, collected transcripts were manually coded based on Garrison et al.'s (2000) CoI coding scheme. Detailed descriptions of each cognitive presence phase and sample quotes are presented in Table 2. For coding analysis of the CS1301 data, two pairs of student research assistants were trained by a researcher experienced in qualitative research. They read the assigned posts and labeled each post with one of the five cognitive presence phases. The inter-rater reliability indicated by the level of agreement between the two coders ranged from 76% to 83%. Likewise, another two pairs of trained student researchers hand coded the CS6601 data under the supervision of the same researcher during the Spring 2021 semester, resulting in inter-rater reliability scores of 94% to 95%.

Table 2
Four Phases of Cognitive Presence and Sample Comments

Cognitive Presence (CP) Index	**CP Phase**	**CP Phase Description**	**Sample Quotes**
0	Non-CP	● Non-cognitive comments ● Socializing comments ● Logistics & technical Q&As	● *Perfect, thanks.* ● *Which chapter is this?*
1	Triggering event	● Expressing confusion ● Disagreement/conflict with prior knowledge ● Clarification questions about a problem	● *I'm so confused by this problem* ● *What do you mean by undersampling, <NAME>?*
2	Exploration of ideas	● Describing/diagnosing a problem ● Sharing hypotheses ● Exploring new ideas or introducing suggestions	● *gah! Still having trouble with the k folds test; it looks like it's breaking something in my confusion matrix* ● *as far as I understood once I re-watched the video, that we're dealing with console-like interfaces to help us focus, examples of those are Pycharm and IDLE, correct?*
3	Integration of ideas	● Giving/proposing someone solutions by	● *In the instructions, it tells you to take the symbol itself … You don't*

		building on other's comments ● Using textbook references or other credible sources to help find solutions	*need to use unicode for this problem.* ● *For the returns in generate_k_folds function, you are supposed return a list of k folds as explained in the function notes.*
4	Resolution of problem or issue	● Confirmation or validation of the proposed solutions ● Elaborating why/how the solution works in details	● *I had a similar issue at first and realized it was because my local test was calling dt.accuracy() inside of the wrong function, and so with the wrong input data.* ● *That means that your input should not require an argument for those attributes and still run. Eg: … is not required as it has a default value, but it can still be changed if an argument is passed in. Thus, it is optional*

Data Analysis

Inferential statistics. To analyze quantitative data, we used IBM SPSS (version 25) to conduct descriptive and correlation analyses to determine associations among cognitive presence scores, final grade scores, and other key variables such as instructor or TA involvement in a discussion thread. Inferential statistics techniques, including the Chi-square test and independent samples t-test, were also used to compare cognitive presence-related trends between sub-groups within and across the courses.

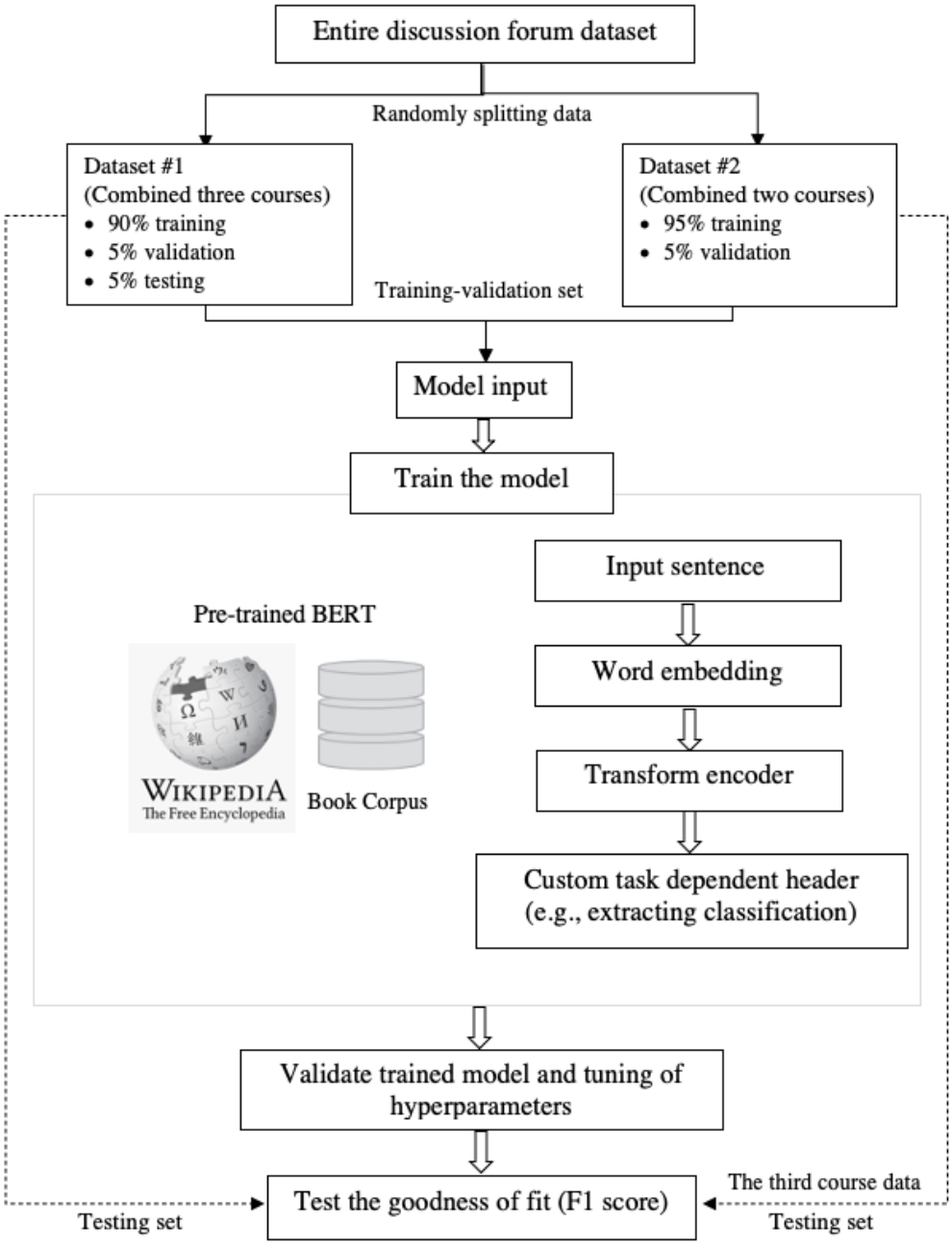
ML-based classification technique. By using the set of manually coded text data that consisted of forum posts and their corresponding cognitive presence scores, we explored an ML approach as a primary technique to automatically identify cognitive presence levels of individual posts generated by participants in discussion forums for online courses. In this case, we were interested in applying a deep neural network technique in which the relationships between input and output elements are predicted based on artificial neural networks that adopt sophisticated and complex modeling algorithms to enable the model to learn and improve predictions over time (Arisoy, 2012).

To effectively train the proposed ML model, we used a transformer-based deep learning model referred to as Bidirectional Encoder Representations from Transformers (BERT), which pre-trains and fine-tunes relevant text data (for an overview see Rogers, Kovaleva, Rumshisky, 2020). BERT was created by Google in 2018 and this is a widely used state-of-the-art technique in many natural language processing tasks to develop language models by learning language representations from unlabeled or uncoded text (Devlin et al., 2019). This model was pre-trained on a large database (around 2,500 million words from Wikipedia and 800 million words from book corpus) and developed by using two different training methods such as Masked Language Model and Next Sentence Prediction. Its size and power make it easily adaptable to novel natural language tasks where there is insufficient data to train a model from scratch.

To fine tune the transformer model, we created training, validation, and test data sets. We tested two different strategies for selecting the subset of data. First, we aggregated all the posts into one large data set and then randomly split it into 90% training, 5% validation, and 5% testing. Next, we clustered posts by course (i.e., CS1301, CS6601) and split them into 95% training and 5% validation. We then took posts from the other course as a test data set to see how

well our model will generalize to posts that were not included in the training data set. Our model used an adaptive SGD algorithm commonly referred to as AdamW, which runs until there is no longer an improvement on the validation data set (Loshchilov & Hutter, 2019). We then selected the model that performed best on the validation data set and compared that to our test dataset. Finally, we computed the F1 score on the test data set as a measure of accuracy, or performance indicator. Figure 1 illustrates the conceptual framework that guided our ML-based classification technique.

Figure 1
Machine-Learning Framework for Classifying Cognitive Presence Phases



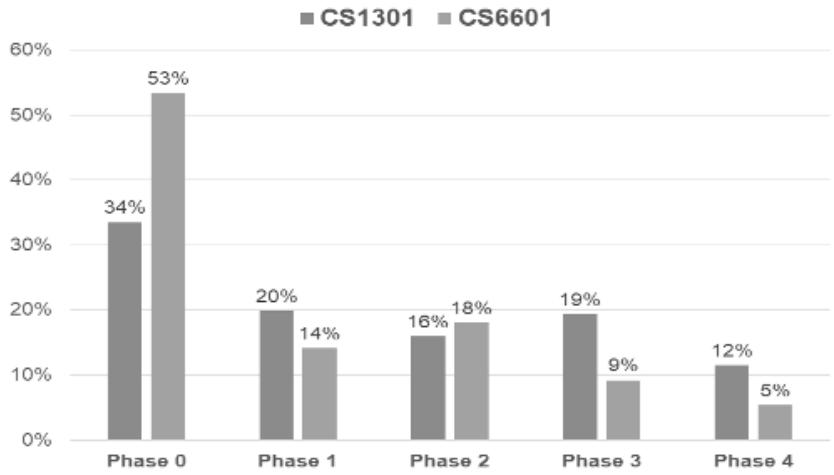
Results

Progression of Cognitive Presence Phases in Online Discussion Forums

First, we examined how online students' cognitive presence develops within the course and whether patterns of idea progression differed by course type. A chi-square test showed that the distribution of cognitive presence phases in students' posts statistically differed between the CS1301 and CS6601 course, X^2 (4, N = 1,896) = 108.90, p < .001. For example, the proportion of Phase 0 comments (e.g., logistics, social) were higher in CS6601 (53%) than in CS1301 (34%) (see Figure 2). This might be because the CS6601 data set specifically came from assignment-related discussion boards and therefore students often asked about the assignment logistics (e.g., deadline extension, grade review). Among comments demonstrating cognitive presence (i.e., Phases 1-4), 46% of total comments in CS1301 (n=540) and 45% of the total in CS6601 (n=678) reflected advanced phases such as integration of ideas and resolution of problems, indicating very similar trends. Interestingly, students in CS1301 posted Phase 1 comments more frequently (30% of total cognitive presence posts) than did those in CS6601 (24% of total). This suggests that MOOC students might introduce problems or seek the input of others more actively, compared to graduate students.

Figure 2

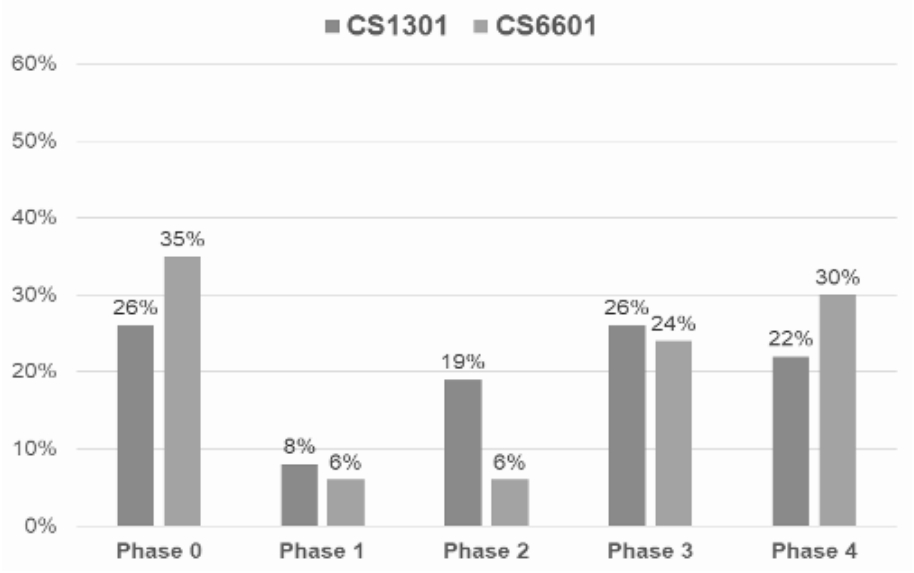
Distribution of Cognitive Presence Phases: CS1301 (MOOC) versus CS6601 (For-Credit Course)



In order to compare progression trends within a specific discussion context, we clustered posts by common discussion thread identifiers and calculated maximum scores of cognitive presence phases for every individual discussion thread. In this case, the maximum cognitive presence score indicated how far participants within a certain discussion thread were able to progress across the four phases of cognitive presence. Then, we compared between the two courses the percentages of threads that generated the maximum cognitive presence score corresponding to each of the four phases (see Figure 3). A chi-square test revealed that the distribution of maximum cognitive presence phases at the thread level statistically differed between the CS1301 and CS6601 course, X^2 (4, N = 505) = 19.13, p < .001. Specifically, compared to CS1301, we observed greater proportions of threads that eventually reached either Phase 0 or Phase 4 in CS6601. This suggests that the graduate-level CS6601 participants frequently stayed in non-cognitive topics but at the same time they were actively engaged with

the problem-solving process to the extent that they advanced to the final phase of cognitive presence. Another notable difference between the two courses was that the percentage of threads that reached Phase 2 was much higher in CS1301 than in CS6601, suggesting that the MOOC discussion forum participants might struggle with going beyond the phase of tackling and exploring problems. This could be also explained by the relatively weak presence of the course instructor or TA as only 36% of the threads in the CS1301 data (total n=350) involved the instructor or TA whereas this figure was 92% in CS6601, pertaining to almost all of the threads that were collected (total n=155).

Figure 3
Distribution of Maximum Cognitive Presence Phases at the Discussion Thread Level: CS1301 versus CS6601



Within each course, we further compared threads that involved the instructor or TAs with threads without their involvement to test whether teaching presence would make any difference in facilitating rapid progression toward the advanced phases, indicated by the degree of changes between the minimum and maximum phases of cognitive presence. According to the t-test results of independent samples, among the CS1301 MOOC students, those who interacted with either the instructor or TA in a discussion thread were likely to show a greater change (M = 1.17) than those who interacted only with their peer students (M = .58), t (338) = 3.99, p < .01. With respect to CS6601, we observed the opposite trend in which students who interacted with the instructor or TAs in a discussion thread tended to exhibit a smaller progression of cognitive presence (M = 1.07) than their peer-only counterparts (M = 1.92), t (150) = -2.15, p < .05. It is possible that students enrolled in a high-stakes online graduate course are poised to deploy critical thinking to solve a problem in the course materials while receiving minimal support from the teaching staff. However, this finding should be viewed with caution given that threads from CS6601 predominantly involved the participation of the instructor or TAs, contributing to an imbalance in the sample sizes between the two groups compared (i.e., 142 threads with teaching presence versus 13 threads without teaching presence).

Online Students' Cognitive Presence and Course Achievement

With respect to the second research question, we examined the relationship between students' levels of cognitive presence and their course achievement. According to correlation analysis results, the maximum cognitive presence scores of individual students had statistically significant, positive, and yet low correlations with final grades in both courses (see Table 3). However, we observed a significant correlation between the students' average cognitive presence scores and their final grades only in CS1301, whereas it was statistically non-significant in CS6601. Based on these findings, we decided to use maximum cognitive presence scores as a primary indicator of the level of cognitive presence that a student was able to achieve in online discussions. Interestingly, there was no significant correlation between the total number of posts that individual students have generated and course grades for CS1301, whereas we observed a significant, positive, and low correlation between the two variables in the CS6601 data. This suggests that, in the MOOC environment, the quantity of participation in discussion forums alone is not meaningfully associated with course achievement. Yet, it is noteworthy that, for both courses, there was a significant, positive, and low to medium correlation between the number of posts and their maximum cognitive presence scores.

Table 3
Bivariate Correlations Among Cognitive Presence Score Variables and Course Grade

	CS1301				CS6601			
	1	2	3	4	1	2	3	4
1. N of Posts	—				—			
2. Max. CP	.26**	—			.39**	—		
3. Avg. CP	.13*	.91**	—		.04	.79**	—	
4. Course Grade	.08	.16**	.14**	—	.21**	.16*	.07	—

Next, we compared the mean course grade scores between students who exhibited different levels of cognitive engagement. In this case, we focused on comparing student subgroups within each course based on how far a student was able to progress through the phases of cognitive presence during online discussions. Within each of the two courses, we calculated the median value of maximum cognitive presence scores among participating students, resulting in a value of 1 for CS1301 and 2 for CS6601. Then, students whose maximum cognitive presence score was either below or corresponding to the median value were assigned to the Low subgroup. Those who had produced a maximum cognitive presence score above the median value were assigned to the High subgroup. That is, in the CS1301 data, students who reached the Phase of 2, 3 or 4 in the discussion forums were categorized as High; while students whose maximum cognitive presence score was either 0 or 1 were categorized as Low. In the CS6601 data, students whose maximum cognitive presence score was either 3 or 4 were categorized as High and those who scored 0, 1 or 2 were categorized as Low.

The independent samples t-test results revealed that the High subgroup was likely to report higher course grades than did the Low subgroup in both courses (see Table 4). In other words, regardless of whether it was a low- or high-stakes course, students who had engaged in higher-order thinking during the collaborative knowledge building process tended to perform better compared to those who had posted only non-cognitive comments or tried to tackle a problem rather at the surface level. The results support the importance of fostering critical thinking in discussion forums to enhance learning outcomes. Interestingly, when the threshold for the subgroup categorization in the CS6601 data was lowered to be equivalent to the CS1301

threshold (i.e., Phase 1), the High and Low subgroups no longer showed a significant difference. This implies that progressing beyond the phase of exploring ideas might have an even stronger impact on the learning outcomes of graduate students.

Table 4
Comparison of Mean Final Course Grades: Independent Samples T-Test Results

	High CP Group			**Low CP Group**			***t*-test**
	n	M	SD	n	M	SD	**(df)**
CS1301	157	49.03	42.23	205	38.29	41.08	-2.44* (360)
CS6601	61	93.97	5.70	115	91.25	6.93	-2.63** (174)

Note. * Indicates $p < .05$. ** indicates $p < .01$.

Applying Machine Learning to Cognitive Presence Identification

For our third research question, we applied ML algorithms to automate the classification of cognitive presence in discussion forum texts. We used a held-out validation data set in which we combined manually coded forum posts collected from both CS1301 and CS6601 and then randomly split the data into training, validation, and test sets. Our pre-trained BERT model was fine-tuned on the training data which accounted for 90% of the entire data set. Eventually, our model achieved a F1 score value of 92.5% on the test data, indicating a high level of accuracy of the model. The F1 score was not only close to our best interrater reliability score (95%) from manual coding but also even higher than the interrater reliability scores that we achieved when coding the CS1301 data. We consider the interrater reliability scores to be our best example of human-level performance on the task, and therefore we are encouraged that our model approached this level of accuracy.

As shown in Figure 4, when compared to the actual coding results, the final model generally performed well in learning to predict both the non-cognitive phase and four phases of cognitive presence. Additionally, as shown in the training curve in Figure 5, we observed that prediction errors, indicated by root mean square error (RMSE), decreased drastically over time as we repeated training sessions. These findings suggest the application of the ML approach to at-scale online learning data such as those data generated from discussion forum posts holds much promise.

Figure 4
Confusion Matrix with Actual versus Predicted CP Phase: Using Combined Data for Training and Testing

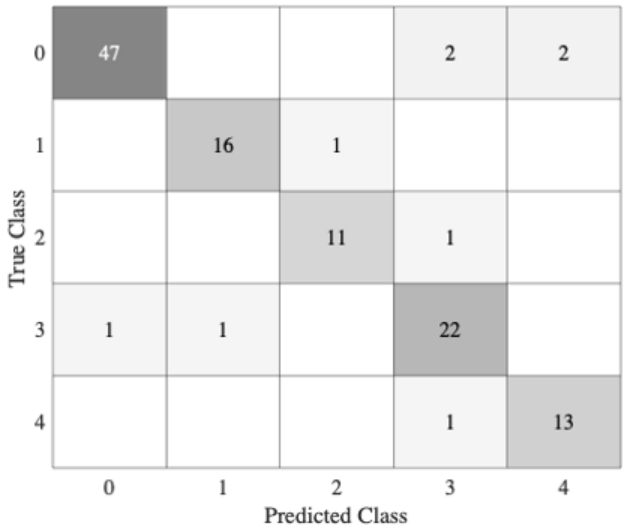
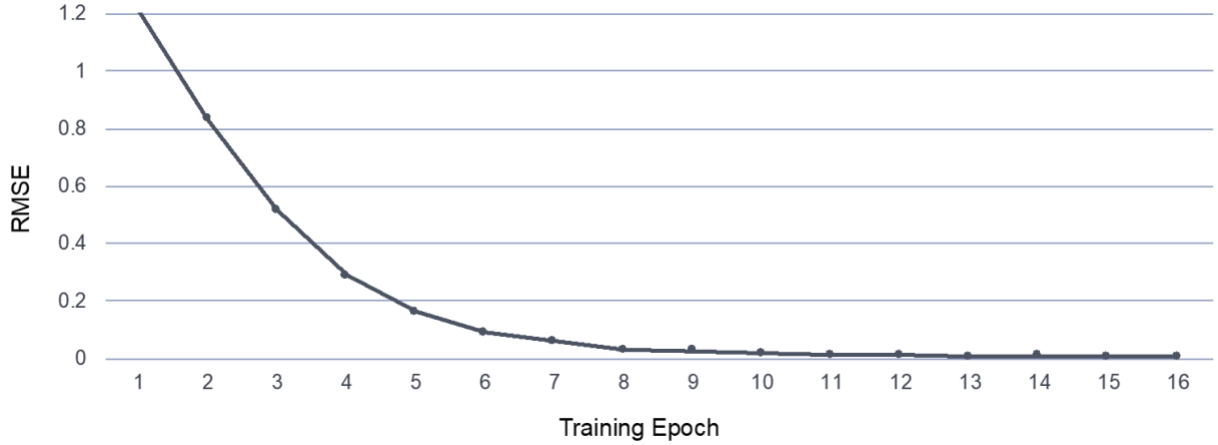


Figure 5
Combined Data Set Training Curve



Note. Training epoch refers to the number of passes of the entire training data set through the machine learning algorithms.

It is notable that we observed much less success in the model performance when we attempted to treat the data from each course separately by using data from one of the two courses to train the model (see Table 5). For example, when the model was trained and tested on the CS1301 data, it achieved a F1 score of only 46.4%. For the CS6601 data, the model performed slightly better than the CS1301 data and yet achieved a much lower F1 score (72.1%) compared to the combined data model (92.5%). Likewise, we observed lower model performances when we combined the data from the two courses for training and then tested the model against the data from a single course (see Table 6). The second procedure resulted in slightly improved accuracy in predicting the cognitive presence phases as indicated by higher F1 scores with 48.9% for CS1301 and 76.6% for CS6601. However, the results clearly suggest that the models in both procedures failed to obtain human-level accuracy in this prediction task.

Table 5
Confusion Matrix with Actual versus Predicted CP Phase: Using Training and Test Data from Specific Course Data

	CS1301 (F1 Score: 0.464)					**CS6601 (F1 Score: 0.721)**				
	Predicted					***Predicted***				
Actual	0	1	2	3	4	0	1	2	3	4
0	12	1	1	0	0	43	0	0	2	0
1	3	7	4	1	0	3	5	3	0	0
2	0	6	1	1	0	2	1	6	1	1
3	1	0	2	0	2	1	0	2	1	0
4	0	0	0	0	1	2	0	0	1	2

Table 6
Confusion Matrix with Actual versus Predicted CP Phase: Using All Data for Training and Using Specific Course Data for Test Data

	CS1301 as Test Data (F1: 0.489)					**CS6601 as Test Data (F1: 0.766)**				
	Predicted					***Predicted***				
Actual	0	1	2	3	4	0	1	2	3	4
0	16	0	1	0	0	31	1	1	1	0
1	3	0	3	1	0	1	6	5	0	0
2	2	0	4	3	0	2	2	10	1	0
3	1	0	3	2	1	1	0	1	6	1
4	0	0	0	1	1	0	0	0	0	4

Overall, our results indicate that a small number of training samples from a single course is not sufficient to fine-tune large, general-purpose language models to the cognitive presence identification task. This is intuitive, as learning to effectively identify cognitive presence requires the ability to generalize across discussion forums with a wide range of interactions and language usage. These results suggest that ML systems for cognitive presence identification should be generalizable to multiple related courses rather than specialized for a single course, since such systems are able to learn more effectively to identify cognitive presence without overfitting to the language of a particular course. Although we only explore this phenomenon in two computer science courses, future work should extend this to more, potentially unrelated, courses to determine the extent to which this is beneficial.

Discussion

Findings from this study contribute to the current literature on cognitive presence in several ways. First, our findings suggest that how students' cognitive presence manifests and progresses may differ by course type and design. As indicated by the relatively high proportion of non-cognitive phase comments posted by graduate students enrolled in CS6601, discussion forums designed to discuss any questions about specific homework or assignments of a course may hinder the opportunity for students to reach higher levels of cognitive presence. Additionally, students' prior knowledge and motivation appeared to be another factor influencing their development of cognitive presence. Our findings indicate that students enrolled in the

CS1301 MOOC tended to focus on generating posts that reflect lower levels of cognitive presence such as those related to triggering events or exploration. This might be due to students participating in discussions with varying degrees of prior knowledge, mostly weak knowledge, of the course topic (i.e., computing in Python). Also, while neither CS1301 and CS6601 required students to engage in discussion as part of the course grading, a very small subset of the CS1301 students contributed to the discussion and participating students tended to generate even fewer posts, compared to the CS6601 students. In order to facilitate progression toward higher levels of cognitive presence, instructors need to consider incorporating the practical inquiry model-based questions (Sadaf & Olesova, 2017), which would allow students to approach a case or course concept by explicitly reflecting on the four levels of cognitive presence (e.g., proposing a solution through synthesis of ideas, applying the solution to a real-world situation). Furthermore, in terms of teaching in MOOC platforms, it is crucial to not only increase students' awareness of the value of contributing to online discussions but also to offer customized resources for students with different levels of background knowledge to help sustain their engagement with critical thinking.

Second, our study explored whether receiving support from either an instructor or TA(s) will have a positive impact on students' collaborative knowledge building process, as measured by the difference between the minimum and maximum cognitive presence score at the discussion thread level. We further examined whether we would observe such a positive impact in other online course environments. It is notable that we observed a relatively stronger impact of the instructor or TA involvement on students situated in the low-stakes MOOC (i.e., CS1301) than those in the high-stakes, for-credit online course (i.e., CS6601). It is possible that MOOC students may benefit more from immediate support from the instructor or TAs, as it may help students sustain engagement with higher-order thinking and advance their knowledge collaboratively with others in discussion. However, our findings capture only a partial snapshot of the CoI model. Previous research has revealed that an instructor's ability to facilitate both teaching and social presence plays a crucial role in enhancing students' cognitive presence (Garrison et al., 2010; Shea & Bidjerano, 2009). Future research will need to expand our current study by addressing how online students' development of cognitive presence can be affected and supported by teacher presence and social presence.

Third, beyond observing how students develop cognitive presence across various types of online courses, our study yields empirical evidence supporting the idea that cognitive presence matters for students' success in both undergraduate-level and graduate-level at-scale learning environments. Our findings are consistent with Sadaf et al.'s (2021) findings that higher levels of cognitive presence are closely associated not only with students' perceived learning but also with their actual final course grades. Moreover, by using the manually coded discussion forum data, our study showed that the extent to which a student is able to progress through the phases of cognitive presence in online discussion (as measured by the maximum cognitive presence score) can be used as a valuable metric to categorize High versus Low cognitive presence subgroups. It is worthwhile to note that the threshold level for identifying the High versus Low subgroup was higher in the CS6601 data than in the CS1301 data, suggesting that, for online graduate students who have advanced domain knowledge and professional experience, it seems more important to be more deeply and cognitively engaged during discussion. Yet, further research is required to replicate and validate our proposed metric in other asynchronous discussion forum contexts.

Fourth, our interdisciplinary approach combines educational psychology and computer science to provide insight into the potential value of the application of the machine-learning approach to the at-scale online learning context in enhancing students' cognitive engagement. Our automated classifier model revealed its robust capability to learn to detect the phases of cognitive presence in discussion forum posts, supporting findings of existing studies (e.g., Hayati et al., 2020; Hu et al., 2020; Neto et al., 2021). Consistent with Neto et al. (2021), we found that the model performance was successful particularly when we used the combined data set for both training and testing. By considering the recommendation from Neto et al., our study made further progress in testing the generalizability of the model by incorporating data sets collected from two different types of online courses (i.e., MOOC and for-credit online course) that cover related subject areas (i.e., computer science). We expect that these findings will provide useful information to online course designers and instructors. For example, our prediction model can be used to create a learning analytics tool designed to benefit students' online learning by enabling instructors to monitor how their students cognitively engage with, and demonstrate progress on, various topics over time in discussion forums. Moreover, based on the automated prediction of students' posts, our model can be implemented as part of instructional design to inform when a teacher or TA should intervene in students' discussion to help build critical thinking and sustain cognitive engagement.

However, our study findings should be interpreted cautiously due to some limitations. For example, we cannot rule out the possibility of sampling bias. In terms of CS1301, only a small subset of the MOOC students participated in discussion forums and these students are likely to be more motivated to learn course concepts than the majority of the enrolled students. Future research will need to examine whether discussion forum participants and non-participants systematically differ in terms of their academic and demographic backgrounds. Also, our findings pose generalizability issues due to relying on specific computer science subjects. In fact, a substantial number of students' posts included computer programming language and code. Accordingly, our proposed automation model, as is, is unlikely to adequately fit data from other discipline areas such as philosophy. Therefore, researchers should continue to investigate the extent to which the ML approach that we adopted will be applicable to course subjects other than computer science. Another limitation of our study is that we were not able to fully account for time-series aspects of the collected data in our statistical analyses due to difficulties with standardizing time zone differences among students participating around the globe. It would be worthwhile to explore whether there are any interesting associations between the development of online students' cognitive presence levels and timing of responses from their teacher, TAs, or peer students (e.g., can students reach higher cognitive presence phases more quickly when they receive support within a certain period?).

Conclusion

Consistent with prior research, our findings suggest that online discussion forums serve as a learning platform where students can actively develop higher-order thinking through the four phases of cognitive presence, whether they are enrolled in an open-access MOOC or a for-credit course. For both courses, discussion participants who engaged with the problem-solving process more deeply tended to achieve better course outcomes, corroborating the crucial role of cognitive presence in facilitating successful online learning. Finally, our exploratory application of ML provides insight into potential solutions to the challenge of measuring and leveraging cognitive presence in large-scale distributed learning environments in higher education.

Furthermore, the initial success of our machine learning approach to cognitive presence classification from forum data supports the design and development of instructional tools and technical interventions which allow instructors to more effectively monitor and support students' learning process at scale.

Declarations
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

The authors received approval from the ethics review board of the Georgia Institute of Technology, USA for this study.

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access*, *7*, 149464-149478. https://doi.org/10.1109/ACCESS.2019.2943351

Amemado, D., & Manca, S. (2017). Learning from decades of online distance education: MOOCs and the Community of Inquiry Framework. *Journal of e-learning and Knowledge Society, 13*(2). https://www.learntechlib.org/p/180225/

An, H., Shin, S, & Lim, K. (2009). The effects of different instructor facilitation approaches on students' interactions during asynchronous online discussions. *Computers & Education, 53*(3), 749-760. https://doi.org/10.1016/j.compedu.2009.04.015

Arisoy, E., Sainath, T. N., Kingsbury, B., & Ramabhadran, B. (2012, June). Deep neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT* (pp. 20-28). Association for Computational Linguistics. https://dl.acm.org/doi/abs/10.5555/2390940.2390943

Askeroth, J.H., & Richardson, J.C. (2019). Instructor perceptions of quality learning in MOOCs they teach. *Online Learning, 23*(4), 135-159. https://doi.org/doi:10.24059/olj.v23i4.2043

Baglione, S., & Nastanski, M. (2007). The superiority of online discussion: Faculty perceptions. *The Quarterly Review of Distance Education, 8*(2), 139-150.

Baran, E., & Correia, A. (2009). Student-led facilitation strategies in online discussions. *Distance Education, 30*(3), 339-361. https://doi.org/10.1080/01587910903236510

Bliuc, A., Ellis, R., Goodyear, P., & Piggott, L. (2009). Learning through face-to-face and online discussions: Associations between students' conceptions, approaches and academic performance in political science. *British Journal of Educational Technology, 41*(3), 512-524. https://doi.org/10.1111/j.1467-8535.2009.00966.x

Chapman, D., Storberg-Walker, J., & Stone, S. (2008). Hitting reply: A qualitative study to understand student decisions to respond to online discussion postings. *E-Learning and Digital Media, 5*(1), 29-39. https://doi.org/10.2304%2Felea.2008.5.1.29

Chen, B., Chang, Y. H., Ouyang, F., & Zhou, W. (2018). Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education*, *37*, 21-30. https://doi.org/10.1016/j.iheduc.2017.12.002

Cheung, W., Hew, K., & Ng, C. (2008). Toward an understanding of why students contribute in asynchronous online discussions. *Journal of Educational Computing Research*, *38*(1), 29-50. https://doi.org/10.2190%2FEC.38.1.b

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Darabi, A., Arrastia, M., Nelson, D., Cornille, T. and Liang, X. (2011). Cognitive presence in asynchronous online learning: A comparison of four discussion strategies. *Journal of Computer Assisted Learning, 27*(3), 216-227. https://doi.org/10.1111/j.1365-2729.2010.00392.x

deNoyelles, A., Zydney, J., & Chen, B. (2014). Strategies for creating a community of inquiry through online asynchronous discussions. *Journal of Online Learning and Teaching, 10*(1), 153-165.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Galikyan, I., Admiraal, W., & Kester, L. (2021). MOOC discussion forums: The interplay of the cognitive and the social. *Computers & Education*, *165*, 104133. https://doi.org/10.1016/j.compedu.2021.104133

Gao, F., Zhang, T., & Franklin, T. (2013). Designing asynchronous online discussion environments: Recent progress and possible future directions. *British Journal of Educational Technology, 44*(3), 469-483. https://doi.org/10.1111/j.1467-8535.2012.01330

Garrison, D. R., & Akyol, Z. (2015). Toward the development of a metacognition construct for communities of inquiry. *The Internet and Higher Education, 24*, 66–71. https://doi.org/10.1016/j.iheduc.2014.10.001

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15*(1), 7-23. https://doi.org/10.1080/08923640109527071

Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education, 13*(1-2), 5-9. https://doi.org/10.1016/j.iheduc.2009.10.003

Guo, P., Saab, N., Wu, L., & Admiraal, W. (2021). The Community of Inquiry perspective on students' social presence, cognitive presence, and academic performance in online project-based learning. *Journal of Computer Assisted Learning, 37*(5), 1479–1493. https://doi. org/10.1111/jcal.12586

Hayati H., Idrissi M. K., & Bennani S. (2020) Automatic classification for cognitive engagement in online discussion forums: Text mining and machine learning approach. In Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., & Millán, E. (Eds.), *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science, vol 12164* (pp. 114-118). Springer, Cham. https://doi.org/10.1007/978-3-030-52240-7_21

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review, 12*, 45-58. https://doi.org/10.1016/j.edurev.2014.05.001

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis

approach. *Computers & Education, 145*, 103724. https://doi.org/10.1016/j.compedu.2019.103724

Hu, Y., Donald, C., Giacaman, N., & Zhu, Z. (2020, March). Towards automated analysis of cognitive presence in MOOC discussions: A manual classification study. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 135-140). https://doi.org/10.1145/3375462.3375473

Irish, I., Finkelberg, R., Nkemelu, D., Gujrania, S., Padiyath, A., Raman, S., Tailor, C., Arriaga, R., & Starner, T. (2020, August). PARQR: Automatic Post Suggestion in the Piazza Online Forum to Support Degree Seeking Online Masters Students. In *Proceedings of the Seventh ACM Conference on Learning@ Scale* (pp. 125-134). https://doi.org/10.1145/3386527.3405914

Kilis, S., & Yildirim, Z. (2019). Posting patterns of students' social presence, cognitive presence, and teaching presence in online learning. *Online Learning, 23*(2), 179-195. https://doi.org/10.24059/olj.v23i2.1460

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016, April). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 15-24). https://doi.org/10.1145/2883851.2883950

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education—A literature review. In A. Peña-Ayala (Ed.), *Learning analytics: fundaments, applications, and trends* (pp. 1-23), Springer. https://doi.org/10.1007/978-3-319-52977-6_1

Mazzolini, M., & Maddison, S. (2007). When to jump in: The role of the instructor in online discussion forums. *Computers & Education., 49*(2), 193-213. https://doi.org/10.1016/j.compedu.2005.06.011

Nanzi, D., Hamilton, M., & Harland, J. (2012). Evaluating the quality of interaction in asynchronous discussion forums in fully online courses. *Distance Education, 33*(1), 5-30. https://doi.org/10.1080/01587919.2012.667957

Neto, V., Rolim, V., Cavalcanti, A. P., Lins, R. D., Gasevic, D., & Ferreiramello, R. (2021). Automatic Content Analysis of Online Discussions for Cognitive Presence: A Study of the Generalizability across Educational Contexts. *IEEE Transactions on Learning Technologies, 14*(3), 299-312. https://doi.org/10.1109/TLT.2021.3083178

Pelánek, R. (2020, March). Learning analytics challenges: Trade-offs, methodology, scalability. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 554-558). https://doi.org/10.1145/3375462.3375463

Quintana, R. M., Pinto, J. D., & Tan, Y. (2021). What we learned when we compared discussion posts from one MOOC hosted on two platforms. *Online Learning, 25*(4), 7-24. https://doi.org/10.24059/olj.v25i4.2897

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics, 8*, 842-866. https://doi.org/10.1162/tacl_a_00349

Sadaf, A., Kim, S. Y., & Wang, Y. (2021). A comparison of cognitive presence, learning, satisfaction, and academic performance in case-based and non-case-based online discussions. *American Journal of Distance Education, 35*(3), 214-227. https://doi.org/10.1080/08923647.2021.1888667

Shea, P., & Bidjerano, T. (2009). Community of inquiry as a theoretical framework to foster "epistemic engagement" and "cognitive presence" in online education. *Computers & Education, 52*(3), 543- 553. https://doi.org/10.1016/j.compedu.2008.10.007

Sadaf, A., & Olesova, L. (2017). Enhancing cognitive presence in online case discussions with questions based on the practical inquiry model. *American Journal of Distance Education, 31*(1), 56-69. https://doi.org/10.1080/08923647.2017.1267525

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 30-32.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*, 98-110. https://doi.org/10.1016/j.chb.2018.07.027

Zhu, M., Bonk, C. J., & Sari, A.R. (2018). Instructor experiences designing MOOCs in higher education: Pedagogical, resource, and logistical considerations and challenges. *Online Learning*, *22*(4), 203-241. https://doi.org/10.24059/olj.v22i4.1495

Appendix

Table A.1
Summary of Existing Literature in the Application of Machine Learning to Online Learning Research

Example Literature		Study Purpose & Scope of ML Application	Online Learning Setting	Methodology for ML Analysis
Authors	Year			
Kovanović et al.	2016	Explored a set of linguistic features of online discussion messages and tested automation of cognitive presence classification	Online Master's level course in software engineering	Random forest classification
Al- Shabandar et al.	2019	Predicted online student performance/dropout and detected at-risk students based on their motivation trajectories & clickstream behaviors	Undergraduate-level MOOCs with various course topics	Random forest classification, generalized linear model, gradient boosting, neural networks, feature selection
Hew et al.	2020	Predicted student satisfaction with MOOCs using data collected through text mining	Randomly selected MOOCs from *Class Central* course metadata	Gradient boosting
Hayati, Idrissi, & Bennani	2020	Classified students into one of four levels of cognitive engagement (i.e., passive, active, constructive, interactive) based on their cognitive behaviors & social interactions within discussion forums	Online courses in software engineering	Support vector machines -based classifier
Neto et al.	2021	Explored a set of linguistic features of online discussion messages (written in Brazilian Portuguese) that can predict the phases of cognitive presence	Online undergraduate courses in biology & technology	Random forest classification