

# Leveraging Large Language Models to Detect and Summarize Cognitive Presence in Online Discussion Forums

Jeonghyun Lee  
Center for 21<sup>st</sup> Century Universities  
Georgia Institute of Technology  
Atlanta, GA, USA  
jonnalee@gatech.edu

Gayane Grigoryan  
Center for 21<sup>st</sup> Century Universities  
Georgia Institute of Technology  
Atlanta, GA, USA  
ggrigoryan3@gatech.edu

Rushil Desai  
Center for 21<sup>st</sup> Century Universities  
Georgia Institute of Technology  
Atlanta, GA, USA  
rdesai316@gatech.edu

Adrian J Gallard  
Center for 21<sup>st</sup> Century Universities  
Georgia Institute of Technology  
Atlanta, GA, USA  
agallard3@gatech.edu

Meryem Yilmaz Soylu  
Center for 21<sup>st</sup> Century Universities  
Georgia Institute of Technology  
Atlanta, GA, USA  
meryem@gatech.edu

**Abstract**—Asynchronous discussion forums are a key pedagogical feature of MOOCs and online degree programs, yet assessing the quality of student engagement at scale remains a persistent challenge. This paper explores the use of large language models (LLMs) to automatically detect cognitive presence and summarize cognitive engagement in online discussions. Grounded in the Community of Inquiry framework, we examine how LLMs can support instructors in understanding the depth of student thinking across large enrollments. We analyzed over 1,500 discussion posts from an introductory MOOC and a graduate-level online computer science course. Using LLaMA-based models, we implemented classification pipelines to categorize posts by cognitive presence phases (triggering event, exploration, integration, and resolution). We compared single-agent and multi-agent architectures where LLMs cross-validate or critique each other’s output to evaluate model accuracy, reliability, and interpretability. Our best-performing models achieved over 92% agreement with human coders. In addition to classification, we employed LLMs to summarize the cognitive trajectory of entire discussion threads, surfacing patterns in student reasoning and critical thinking. Findings suggest that LLMs can provide scalable, accurate insights into students’ cognitive engagement, with both single- and multi-agent systems showing promise for enhancing model robustness. We discuss the pedagogical implications of integrating these tools into digital learning environments and outline key considerations for model generalizability. This study contributes to the growing field of AI-enhanced learning analytics, offering a practical framework for automated feedback in large-scale online education.

**Keywords**—MOOC, online course, discussion forums, large language model, generative AI

## I. INTRODUCTION

Asynchronous discussion forums offer a unique learning opportunity for online students to interact and construct knowledge with others. Previous research shows that participation in online discussions at a deep level through critical thinking, reflecting, and refining meaning has been found to enhance student learning [1], [2]. According to the community of inquiry (CoI) theory [3], a widely adopted framework to guide research in learning in collaborative online learning environments, meaningful knowledge construction can be fostered through the critical dimensions of teaching, social, and cognitive presence [2], [4]. However, assessing cognitive presence poses persistent challenges in

asynchronous discussion forums, especially within MOOCs and large-scale online programs where instructors cannot feasibly read thousands of posts. As a result, opportunities for timely feedback and intervention are often missed, which can negatively impact students’ engagement and overall quality of learning. In this study, we explore how large language models (LLMs) can be used to automatically summarize and detect cognitive presence in online discussions at scale.

### A. The Role of Cognitive Presence in Discussion Forums

Cognitive presence, a critical dimension of the CoI framework, refers to the extent to which students engage in reflective thinking and critical inquiry [3]. Cognitive presence can be manifested in four developing 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). 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 [4], [5].

Empirical studies suggest that when students effectively engage in higher phases of cognitive presence such as integration and resolution, they are more likely to apply course concepts meaningfully and transfer knowledge to novel contexts [5], [6]. Additionally, online discussions characterized by high cognitive presence have been shown to foster active collaboration and promote peer-to-peer learning and collective knowledge construction [7], [8]. This is especially crucial in large-scale online learning environments like MOOCs in which direct instructor engagement is often limited [9]. However, measuring and supporting cognitive presence remains an ongoing challenge for educators and researchers who are seeking to improve the quality of online learning in large-enrollment courses.

### B. Measurement of Cognitive Presence

Traditionally, cognitive presence has been measured through qualitative coding methods, which requires extensive manual analysis of discussion posts data. Such method becomes impractical as course sizes increase. To address this challenge, automated methods, including machine learning (ML) and natural language processing (NLP), have increasingly been employed to analyze students’ cognitive

engagement due to their potential scalability [10], [11], [12], [13]. Researchers have explored supervised classification models that use ML algorithms to categorize discussion posts according to the four cognitive presence phases. For example, random forest classifiers have been frequently used to categorize cognitive presence phases based on linguistic features from course discussion forums [13], [14], [15].

Although these studies collectively highlight ML's potential to provide scalable insights into students' cognitive engagement and learning outcomes in digital educational contexts, results from traditional NLP models, which relied on using dictionary-based features, have shown varied performance and issues of overfitting and low generalizability. For example, Kovanović et al. [16] introduced a Support Vector Machine model that incorporated both n-gram and structural features, improving the accuracy to 58.4%. Building on this, Waters et al. [17] demonstrated that structural features play a key role by applying a Conditional Random Field model, which further increased performance to 64.2% accuracy. Kovanović et al. [18] further considered Random Forest classifier and consider class imbalance techniques for discussion forum classification, and the accuracy achieved with these transformations are 70.3%.

### C. Application and Potential of LLMs in Online Learning Data

To address the limitations of traditional ML approaches for detecting cognitive presence, such as the need for extensive manual feature engineering and the lack of contextual interpretability, researchers have recently turned to transformer-based LLMs. These models including Bidirectional Encoder Representation Transformers (BERT) offer advanced natural language processing capabilities that allow for deeper and more flexible discourse analysis [10], [19], [20]. For example, Authors [10] leveraged BERT to classify cognitive presence with high accuracy with 92.5% across MOOC and graduate-level online discussions. Similarly, Hu et al., [20] developed fine-tuned multi-label BERT classifier, which demonstrated strong performance in classifying cognitive presence in MOOCs. However, such models often require task-specific tuning and do not provide transparent rationale for classifications.

Guided by the LLM-Assisted Content Analysis (LACA) framework [21], Catellanos-Reyes [22] leveraged OpenAI's GPT models to conduct automated content analysis based on the Practical Inquiry Model. By designing an AI-adapted CP codebook and testing a range of prompting techniques (e.g., one-shot, few-shot, chain-of-thought), the study demonstrated that LLMs could reliably classify student posts, particularly in the more abstract "Integration" phase of cognitive presence. Their findings highlight the potential of LLMs for cognitive presence classification and emphasize LACA's cost-efficiency in educational research. However, the authors also noted the need for considerable data literacy among researchers and educators to deploy LLM tools effectively.

### D. The Current Study

Despite the promise of commercial LLMs such as GPT-4, widespread adoption in educational settings remains limited by their proprietary nature, high cost, and restricted transparency. As a result, open-source alternatives like the LLaMA family have emerged as attractive options for research and practice [23], [24]. Our study builds on this momentum by testing the extent to which the LLaMA-based

models can achieve comparable accuracy in classifying cognitive presence in online discussion posts.

Guided by the CoI framework, our exploratory study seeks to apply open-source AI models to assess students' cognitive presence within discussion forums of large-enrollment online courses. Specifically, we investigate the effectiveness of different LLMs and fine-tuning techniques in accurately identifying the presence and development of critical thinking. Additionally, we explore the capacity of these models to offer contextualized insights that can support educators in enhancing practices in online learning and teaching.

## II. METHOD

### A. Participants and Settings

Participants were enrolled in one of two online courses: an introductory computer programming MOOC (CS1301) and a graduate-level artificial intelligence course (CS6601). Both courses were offered by the same technology-focused public university in the U.S. and were taught by different instructors. CS1301 is a free, open-access course on the edX platform with no programming prerequisites and only basic arithmetic and high school algebra are recommended. This low-stakes course typically enrolls thousands of learners; for instance, we observed nearly 45,000 students registered during in Fall 2017. On the other hand, CS6601 is a high-stakes, for-credit course within the Online Master's in Computer Science program, and it is a much smaller class size than CS1301 (e.g., 796 students in Spring 2020). It requires prior knowledge of college-level mathematical concepts and computer programming and algorithms. Students in both courses were invited to participate voluntarily in the discussion forums, and their participation was not counted toward the final grade.

### B. Data Sources

The data source included a total of 505 randomly selected discussion threads collected from two online courses: 350 threads from the CS1301 MOOC, collected during Fall 2017 and Fall 2018 via edX and 155 threads from the CS6601 graduate-level course, collected during Spring 2020 using the Piazza discussion forum. In total, these threads comprised 2,341 individual discussion posts, with 848 from CS1301 and 1,493 from CS6601. All discussion posts were anonymized prior to data analysis.

### C. Measures and Procedures

All discussion posts were manually coded by researchers based on the four cognitive presence phases defined by the CoI framework [3] (see Table I). Two pairs of trained student research assistants conducted the coding analysis, following instructions and guidance provided by one of the authors. The coding process resulted in inter-rater reliability scores of up to 95%, indicating a high level of agreement among coders.

To prepare the dataset for modeling, the discussion threads data were split into training (80%), validation (10%), and test (10%) subsets. The data were stratified by course and cognitive presence labels to maintain distributional balance across the subsets. The training set included an additional explanation field that described the rationale behind each assigned label. This field was used to support the fine-tuning of LLMs by providing contextual guidance for supervised learning during the analysis process.

TABLE I. COGNITIVE PRESENCE PHASES AND SAMPLE QUOTES

Cognitive Presence Phase	Description (Sample Quotes)
Non-cognitive	Socializing comments (e.g., <i>Thank you!</i> ) Logistics, deadlines for exams or assignments, technical Q&As (e.g., <i>can i download the smartbook?</i> )
Triggering Event	Disagreement/conflict with prior knowledge (e.g., <i>I'm so confused by this problem</i> ) Confusion or need for clarification about a problem (e.g., <i>What do you mean by undersampling?</i> )
Exploration of Ideas	Describing/diagnosing a problem (e.g., <i>gah! Still having trouble with the k folds test; it looks like it's breaking something in my confusion matrix</i> ) Sharing hypotheses (e.g., <i>Are you printing the result instead of returning it?</i> )
Integration of Ideas	Citing textbook references or other credible sources (e.g., <i>In the instructions, it tells you to take the symbol itself from the instructions. You don't need to use unicode for this problem.</i> ) Proposing solutions (e.g., <i>Try deleting your test lines from 45 onwards. If that fails, try resetting.</i> )
Resolution of Problem	Confirmation or validation of ideas (e.g., <i>Correct there are some nuances that can only be taken into account by actually implementing them</i> ) Elaboration on the proposed solution with explanation of why/how the solution works (e.g., <i>You're on the right track, ... you shouldn't print within the function ... you will have to add them (in the right order) to your result string. Also, you should declare your return string before the loop start and then add to it.</i> )

#### D. Data Analysis

*a) Pre-processing vis Recursive Summarization:* To address context length limitations and reduce irrelevant content, we employed a summarization preprocessing step. This approach preserves cognitively relevant discourse signals, such as reasoning steps and peer knowledge building, while reducing thread length for LLM processing. Each discussion thread was summarized before classification to distill the cognitively relevant content. We adopted a MapReduce summarization strategy, well-suited for handling lengthy inputs [25]. Initially introduced by LangChain, the MapReduce method involves two steps: the map step and the reduce step. In the map step, the discussion thread was split into token-bounded segments. Each segment was summarized independently using an LLM. In the reduce step, the resulting segment-level summaries were merged and recursively summarized until the final output fit within the target context window. This recursive collapsing approach enabled summarization of arbitrarily long discussion threads, often including extensive code snippets or error traces, while preserving key signals of student reasoning and engagement (see Table II for an example).

TABLE II. SAMPLE SUMMARY OUTPUT

Summary Output	Note
The conversation revolves around the implementation of a machine learning algorithm, specifically the decision tree algorithm. The student asks about the splitting process, and the TA explains the step-by-step process of splitting the data column by column, starting with the first column, and calculating the Gini gain for each split. The student is interested in finding the column that gives the highest Gini gain. Additionally, the conversation touches on the logic for handling leaf nodes in the decision tree, clarifying that even if a split is made on a particular attribute value, the corresponding row should not be removed from the tree.	The original thread included detailed decision tree splitting calculations, but the summary excludes them, while capturing all cognitive elements.

*b) Cognitive Presence Classification:* The core task of this study was to automatically detect and classify cognitive presence in online discussion posts, as defined by the CoI framework. We approached this as two separate classification tasks. The first task is binary classification to determine whether a post contains cognitive presence or is non-cognitive. The second task is multi-label classification to identify one or more of the cognitive presence phases.

*c) Few-Shot Prompting with RAG:* Given the diversity of discussion styles and topics, we used Retrieval-Augmented Generation (RAG) to dynamically select relevant few-shot examples [26]. We manually labeled a subset of discussions, each with a cognitive presence label and an explanation justifying the label. These labeled examples were embedded and indexed. At inference time, the system retrieved the top-k similar examples ( $k = 3$ ) based on semantic similarity and included them in the prompt. Including explanations significantly improved model accuracy and interpretability by anchoring predictions in grounded reasoning.

*d) LLM Architecture:* We tested multiple variants of the LLaMA family of models, including: LLaMA 2 (7B, 13B), LLaMA 3 (8B, 70B), LLaMA 3.1 (8B, 70B), and LLaMA 3.3 (70B). Our single-agent architecture serves as a baseline for both binary and multilabel cognitive presence classification tasks. In this setup, a single LLM instance is responsible for producing predictions and justifications based on a prompt constructed using dynamic few-shot examples. In regard to the prompt design, the model receives a prompt that includes the input discussion (or its summary), along with  $k = 3$  retrieved examples drawn from a manually annotated dataset. Each example includes both the label(s) and a natural language explanation. This dynamic prompt is generated via RAG to ensure contextual relevance and diversity across different discussions.

To improve robustness and reduce model hallucination or overconfidence, we implemented a generator-reflector multi-agent framework based on reflective collaboration between two LLM instances [27]. The multi-agent setup consists of two roles: generator and reflector. The generator is identical to the single-agent model and produces an initial prediction and explanation based on the RAG-augmented prompt. On the other hand, the reflector assumes the role of a critical peer or instructor. It reviews the generator's output and offers constructive critique, correction, or affirmation. In this framework, both the generator and reflector engage in a fixed-loop interaction to ensure computational efficiency and reliable assessment. After each iteration, the generator revises its output in response to the reflector's critique. This process continues until the reflector agrees with the output or until a maximum of three iterations is reached.

### III. RESULTS

We evaluated the performance of both single-agent and multi-agent architectures on binary and multilabel cognitive presence classification tasks. Model performance was measured using weighted average F1-score, computed from classification outputs against human-coded ground truth.

*a) Single-Agent Classification Performance:* For both the binary and multilabel tasks, we observed a consistent trend. Newer and larger models outperformed older and smaller ones, with accuracy and explanation quality

improving notably with scale. This performance trend indicates that improvements in model architecture and pre-training (e.g., between LLaMA 2 and LLaMA 3.3) have a measurable impact on the model’s ability to accurately identify and explain cognitive presence. The best-performing model results are reported in Table III.

TABLE III. SINGLE-AGENT CLASSIFICATION PERFORMANCE RESULTS

Class	Model Performance Metrics			
	Precision	Recall	F1-Score	Support
<i>Binary Classification (Llama 3.1 70B)</i>				
True	0.86	0.86	0.86	29
False	0.94	0.94	0.94	69
Accuracy			0.92	98
<i>Multi-Label Classification (Llama 3.3 70B)</i>				
Triggering	0.48	0.93	0.64	30
Explorative	0.60	0.83	0.70	36
Integrating	0.75	0.71	0.73	38
Resolution	0.50	0.56	0.53	25
Accuracy (Micro Avg)			0.66	129

The binary classification task involved identifying whether a given discussion thread demonstrated evidence of cognitive presence (Table IV). The model achieved a weighted average F1-score of 0.92. The model performed particularly well on the False class (i.e., non-cognitive presence), with precision, recall, and F1-score all at 0.94, likely due in part to its larger representation in the test set ( $n = 69$ ). While the True class (i.e., cognitive presence), which has fewer instances ( $n = 29$ ), shows slightly lower but still robust metrics (precision, recall, and F1-score all at 0.86), the classifier maintains balanced performance across both categories. These results suggest that the model is both accurate and reliable for distinguishing between the two classes, with minimal evidence of class imbalance effects or systemic bias in prediction.

TABLE IV. SAMPLE BINARY CLASSIFICATION OUTPUT

<b>Model Input (CS6601)</b>	The student is struggling with a Gibbs sampling issue in their coursework, specifically with the calculation of $P(A)$ , $P(B)$ , and $P(C)$ using the 14.12 equation. They provided a link to their submission on Gradescope and asked the instructor to review their work. The TA responded by acknowledging the issue and indicating that it was resolved.
<b>Model Output</b>	<code>{"present": "true", "explanation": "The text shows cognitive presence through a \"Trigger Event\" and \"Integrating Solution.\" The student encounters challenges with Gibbs sampling calculations and seeks help by sharing their submission. The TA acknowledges the issue and confirms its resolution, reflecting a collaborative effort that enhances the student's understanding."}</code>

The multilabel classification model achieved a weighted average F1-score of 0.66, with varying levels of model performance across the four cognitive presence categories (Table V). Notably, the model achieved high recall for the Triggering Event category (0.93), indicating its ability to identify most true instances, but at the cost of low precision

(0.48), suggesting frequent false positives. In contrast, the Exploring Problem and Integrating Solutions categories exhibited relatively balanced performance, with F1-scores of 0.70 and 0.73, respectively, reflecting the model’s capacity to both correctly identify and accurately predict these cognitive processes. The Resolution category proved the most challenging to classify, with the lowest F1-score (0.53), highlighting potential ambiguity or insufficient signal in the input features.

TABLE V. SAMPLE BINARY CLASSIFICATION OUTPUT

<b>Model Input (CS1301)</b>	The discussion revolves around a student's code submission, which functions correctly in Idle but fails in the grader. The student initially attributes this discrepancy to an error message for invalid values. Another student identifies that test code at the end of their file might have affected the grader, which was resolved by removing it. The conversation highlights the importance of case sensitivity and ensuring test code doesn't interfere with grading.
<b>Model Output</b>	<code>{"triggering": "true", "explorative": "true", "integrating": "true", "resolution": "true", "explanation": "The text demonstrates cognitive presence through a \"Trigger Event\" where the student encounters a discrepancy between their local environment and the grader. The \"Explorative Discussion\" is evident as the student investigates the issue, with another student suggesting a potential solution related to case sensitivity. The \"Integrating Solution\" is seen when the student applies the suggestion and confirms the code works. Finally, the \"Resolution\" is achieved when another student points out that test code at the end of the file might have been interfering with the grader, and this issue is resolved by removing the test code. The conversation actively engages with the academic content, promoting deeper understanding and problem-solving."}</code>

The overall results suggest that the model is effective at capturing cognitive engagement, particularly in more frequent categories. However, it requires additional refinement to improve class discrimination, especially for less distinct categories such as Resolution.

*b) Multi-agent (Reflection-Based) Classification Performance:* We evaluated the multi-agent generator-reflector architecture exclusively on the binary classification task. This setup achieved a weighted average F1-score of 0.91, which is close to the best single-agent score (0.92 from LLaMA 3.3 70B) (see Table VI). The "False" class (i.e., posts not exhibiting cognitive presence) was detected with higher precision (0.95) and F1-score (0.93) compared to the "True" class (i.e., posts exhibiting cognitive presence), which had a precision of 0.81 and F1-score of 0.85. The recall was relatively balanced across both classes, at 0.90 for "True" and 0.91 for "False", indicating the model's consistent ability to retrieve relevant examples from both categories. These results suggest that while the model effectively identified posts that lacked cognitive presence, it adopted a more conservative approach when labeling posts as cognitively present likely to reduce the risk of false positives.

TABLE VI. MULTI-AGENT CLASSIFICATION PERFORMANCE RESULTS

Class	Model Performance Metrics			
	Precision	Recall	F1-Score	Support
<i>Binary Classification (Llama 3 70B)</i>				
True	0.81	0.90	0.85	29
False	0.95	0.91	0.93	69
Accuracy			0.91	98

As illustrated in Figure 1, the reflective loop facilitated useful refinements in cases of ambiguity and enabled better insight into the model's decision-making process, supporting more transparent and pedagogically interpretable output. While the overall F1-score was comparable, we anticipate that the multi-agent framework will be especially valuable for complex multi-label classification, which remains an area for future work.

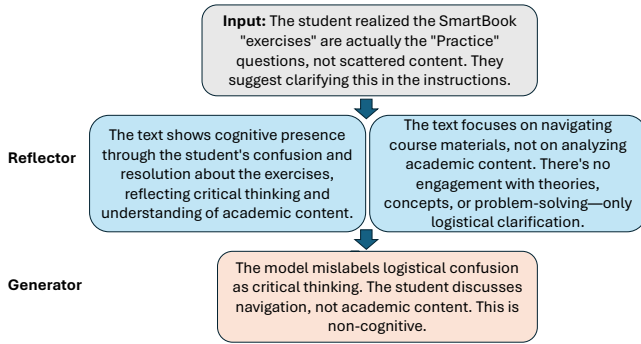


Fig. 1. Example Multi-Agent Output

#### IV. DISCUSSION

This study demonstrates the potential of LLMs in detecting and classifying cognitive presence in discussion forums, particularly within large-scale online learning environments such as MOOCs and online graduate programs. By exploring the effectiveness of both single-agent and multi-agent LLM architectures, we shed light on the applicability of AI tools in assessing and supporting learning processes that were previously difficult to observe at scale [12], [20], [28]. Specifically, our findings highlight three key contributions.

First, we showcase that LLMs can offer a scalable solution to understanding how students engage cognitively in online discussions by efficiently and accurately processing and classifying complex text data. Our approach reduces the time and labor associated with analyzing discussion threads, allowing for near real-time feedback and timely pedagogical adjustments. This provides educators and researchers with a powerful tool for assessing learning engagement in large cohorts, where manual content analysis is often impractical.

Also, the results demonstrate that LLMs can summarize and identify patterns of cognitive presence in a way that helps educators make sense of how students build knowledge and engage in critical thinking through collaborative discourse. This has important implications for course design and instructional interventions, especially in asynchronous online environments where instructor-student interaction is more limited [5], [6]. Additionally, this opens up new possibilities for adaptive learning systems that can respond dynamically to students' evolving needs and engagement levels.

Next, this study extends the growing body of research on AI in education by demonstrating that open-source LLMs, particularly the LLaMA family, can effectively classify cognitive presence in online discussion forums. We systematically evaluated multiple LLaMA variants and found a clear performance trend: newer and larger models, such as LLaMA 3.3 70B, achieved superior accuracy and explanation quality. This model reached a F1-score of 0.92 for binary classification and 0.66 for multilabel classification. Importantly, our results show that open-source LLMs can serve as affordable, scalable tools for educators and researchers. By relying on freely available models rather than commercial APIs, we lower the technical and financial barriers to applying AI in real-world educational settings. This work reveals the practical viability of open-source models for educational text classification tasks.

Furthermore, we compared a multi-agent reflection-based framework against single-agent models and found that its F1-score for the binary classification task was slightly lower (0.91). This raises questions about the specific conditions under which a multi-agent approach might offer distinct advantages. While we observed that multi-agent models could potentially excel in more ambiguous or nuanced phases of cognitive presence, the results suggest the need for further research to determine the most effective deployment scenarios for each architecture.

Despite these contributions, our findings should be interpreted with caution due to several limitations. Most notably, our multi-label classification of the four phases of cognitive presence achieved an accuracy score of only 66%, which is lower than expected. Prior research has shown that detecting the higher phases of cognitive presence (i.e., integration and resolution) is inherently difficult due to their abstract, complex, and often implicit nature [4], [22]. This limitation reflects the broader challenge of aligning AI-generated classifications with the nuanced interpretations made by human coders in contexts such as critical thinking and knowledge construction. These findings suggest a need for further refinement in prompt engineering strategies and for developing clearer and more operationalized definitions of cognitive presence phases to enhance model reliability and theoretical alignment.

Another limitation of this study lies in the use of data combined from two distinct online courses, limiting generalizability of our findings. Moreover, such contextual differences likely influenced how cognitive presence was expressed and perceived, making it more difficult for the model to learn consistent patterns across datasets. These inconsistencies may have contributed to reduced classification accuracy, particularly in distinguishing between subtle phases of cognitive engagement. Future research should explore the impact of course-specific features on model performance and consider whether training separate models for individual courses or incorporating course-level context during model fine-tuning can improve robustness.

#### V. CONCLUSION

In conclusion, this study contributes to the advancement of digital education and MOOC learning research by illustrating how open-source LLMs can effectively support the large-scale analysis of online discussion forum posts. By evaluating

both single-agent and multi-agent LLM architectures, we demonstrate that these tools are capable of classifying and interpreting students' cognitive presence or critical thinking with increasing accuracy. Our findings underscore the practical value of affordable AI solutions for instructional support in large, asynchronous learning environments. Yet, while LLMs offer timely insights into student thinking processes, this work points to the ongoing need for improving their interpretive capabilities and adapting them to diverse learning contexts, which becomes a critical direction for future research.

## REFERENCES

- [1] A. M. Bliuc, R. A. Ellis, P. Goodyear, and L. Piggott, "A blended learning approach to teaching foreign policy: Student experiences of learning through face-to-face and online discussion and their relationship to academic performance," *Computers & Education*, vol. 56, no. 3, pp. 856–864, 2011. <https://doi.org/10.1016/j.compedu.2010.10.027>
- [2] I. Galikyan, W. Admiraal, and L. Kester, "MOOC discussion forums: The interplay of the cognitive and the social," *Computers & Education*, vol. 165, 104133, 2021. <https://doi.org/10.1016/j.compedu.2021.104133>
- [3] D. R. Garrison, T. Anderson, and W. Archer, "Critical thinking, cognitive presence, and computer conferencing in distance education," *Am. J. Distance Educ.*, vol. 15, no. 1, pp. 7–23, 2001. <https://doi.org/10.1080/08923640109527071>
- [4] D. R. Garrison, T. Anderson, and W. Archer, "The first decade of the community of inquiry framework: A retrospective," *Internet High. Educ.*, vol. 13, no. 1–2, pp. 5–9, 2010. <https://doi.org/10.1016/j.iheduc.2009.10.003>
- [5] A. Sadaf and L. Olesova, "Enhancing cognitive presence in online case discussions with questions based on the practical inquiry model," *Am. J. Distance Educ.*, vol. 31, no. 1, pp. 56–69, 2017. <https://doi.org/10.1080/08923647.2017.1267525>
- [6] Z. Akyol and D. R. Garrison, "Assessing metacognition in an online community of inquiry," *Internet High. Educ.*, vol. 14, no. 3, pp. 183–190, 2011. <https://doi.org/10.1016/j.iheduc.2011.01.005>
- [7] R. L. Moore and C. N. Miller, "Fostering cognitive presence in online courses: A systematic review (2008–2020)," *Online Learning*, vol. 26, no. 1, pp. 130–149, 2022.
- [8] P. Shea and T. Bidjerano, "Community of inquiry as a theoretical framework to foster 'epistemic engagement' and 'cognitive presence' in online education," *Computers & Education*, vol. 52, no. 3, pp. 543–553, 2009. <https://doi.org/10.1016/j.compedu.2008.10.007>
- [9] Y. Chen, Q. Gao, Q. Yuan, and Y. Tang, "Discovering MOOC learner motivation and its moderating role," *Behav. Inf. Technol.*, vol. 39, no. 12, pp. 1257–1275, 2020. <https://doi.org/10.1080/0144929X.2019.1661520>
- [10] Authors, 2022.
- [11] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Computers & Education*, vol. 145, 103724, 2020. <https://doi.org/10.1016/j.compedu.2019.103724>
- [12] V. Kovanović et al., "Examining communities of inquiry in massive open online courses: The role of study strategies," *Internet High. Educ.*, vol. 40, pp. 20–43, 2019. <https://doi.org/10.1016/j.iheduc.2018.09.001>
- [13] V. Dornauer et al., "Automatic classification of online discussions and other learning traces to detect cognitive presence," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 395–415, 2024. <https://doi.org/10.1007/s40593-023-00335-4>
- [14] V. Kovanović et al., "Understanding the relationship between technology use and cognitive presence in MOOCs," in *Proc. 7th Int. Learn. Analytics & Knowl. Conf.*, 2017, pp. 582–583.
- [15] V. Neto et al., "Automatic content analysis of online discussions for cognitive presence: A study of the generalizability across educational contexts," *IEEE Trans. Learn. Technol.*, vol. 14, no. 3, pp. 299–312, 2021. <https://doi.org/10.1109/TLT.2021.3083178>
- [16] V. Kovanović et al., "Automated cognitive presence detection in online discussion transcripts," in *Proc. Workshops at 4th Int. Conf. Learn. Analytics & Knowl.*, 2014.
- [17] Z. Waters et al., "Structure matters: Adoption of structured classification approach in the context of cognitive presence classification," *Lect. Notes Comput. Sci.*, vol. 9460, pp. 227–238, 2015. [https://doi.org/10.1007/978-3-319-28940-3\\_18](https://doi.org/10.1007/978-3-319-28940-3_18)
- [18] V. Kovanović et al., "Towards automated content analysis of discussion transcripts: A cognitive presence case," in *Proc. 6th Int. Conf. Learn. Analytics & Knowl.*, 2016, pp. 15–24. <https://doi.org/10.1145/2883851.2883950>
- [19] E. Eryilmaz, B. Thoms, and Z. Ahmed, "Theory-guided multiclass text classification in online academic discussions," *J. Comput. Inf. Syst.*, pp. 1–12, 2024. <https://doi.org/10.1080/08874417.2024.2371435>
- [20] Y. Hu, C. Donald, and N. Giacaman, "A revised application of cognitive presence automatic classifiers for MOOCs: A new set of indicators revealed?," *Int. J. Educ. Technol. High. Educ.*, vol. 19, no. 1, 48, 2022.
- [21] R. Chew et al., "LLM-assisted content analysis: Using large language models to support deductive coding," *arXiv preprint arXiv:2306.14924*, 2023.
- [22] D. Castellanos-Reyes, L. Olesova, and A. Sadaf, "Transforming online learning research: Leveraging GPT large language models for automated content analysis of cognitive presence," *Internet High. Educ.*, vol. 58, 101001, 2025.
- [23] M. P. C. Lin, D. Chang, S. Hall, and G. Jhaji, "Preliminary systematic review of open-source large language models in education," in *Proc. Int. Conf. Intell. Tutor. Syst.*, Cham, Switzerland: Springer Nature, 2024, pp. 68–77.
- [24] C. Koutchme, N. Dainese, S. Sarsa, A. Hellas, J. Leinonen, and P. Denny, "Open source language models can provide feedback: Evaluating LLMs' ability to help students using GPT-4-as-a-judge," in *Proc. 2024 Innov. Technol. Comput. Sci. Educ. V. 1*, 2024, pp. 52–58.
- [25] H. Chase, *Langchain*, GitHub, 2022. [Online]. Available: <https://github.com/langchain-ai/langchain>
- [26] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [27] A. Madaan et al., "Self-refine: Iterative refinement with self-feedback," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 46534–46594, 2023.
- [28] A. F. Wise, Y. Cui, W. Jin, and J. Vytasek, "Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling," *Internet High. Educ.*, vol. 32, pp. 11–28, 2017. <https://doi.org/10.1016/j.iheduc.2016.08.001>