Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
Paper ID #48207

# BOARD # 288: NSF: IUSE Harnessing Language Models to Predict and Enhance STEM Engagement Using Non-Cognitive Experiential Data

**Ahatsham Hayat, University of Nebraska - Lincoln**
**Bilal Khan, Lehigh University**
**Mohammad Rashedul Hasan, University of Nebraska - Lincoln**

# NSF: IUSE Harnessing Language Models to Predict and Enhance STEM Engagement Using Non-Cognitive Experiential Data

## Abstract

This research explores the use of pre-trained large language models (LLMs) to predict weekly lecture-based engagement of college STEM students based on longitudinal experiential data. We leverage non-cognitive attributes, such as emotional responses, and socio-economic background information to forecast engagement patterns. To address data limitations, we employ a contextual data enrichment method. Experiments with BERT (encoder-only) and Llama (decoder-only) models demonstrate that BERT achieves higher accuracy, particularly with non-cognitive data, while both models improve with background data integration. These findings highlight LLMs' potential to enable data-driven interventions in STEM education by predicting student engagement.

## Introduction

Recent advancements in artificial intelligence (AI), particularly Transformer-based large language models (LLMs) [1], have demonstrated remarkable performance across various downstream natural language processing (NLP) tasks by implicitly encoding vast amounts of general knowledge within their parameters [2, 3] and exhibiting advanced linguistic capabilities [4]. These models, known for their versatility across domains [5, 2], hold considerable promise for applications in educational settings. While previous work has explored their use in forecasting end-of-semester academic performance from longitudinal experiential (LE) data [6, 7], which systematically captures students' cognitive and behavioral experiences over time, the application of AI models to predict student engagement on a regular basis remains underexplored.

Predicting student engagement is crucial for identifying key indicators of academic success, addressing challenges, and enabling timely interventions [8, 9]. Lecture-based engagement, which encompasses student attentiveness, participation, and cognitive involvement, is strongly linked to academic performance and long-term retention in science, technology, engineering, and mathematics (STEM) fields. LE data, typically gathered through self-reports, provides insights into students' perceptions and emotions within educational settings [10]. However, the subjective nature of self-reported data and its temporal variability introduce challenges in accurately capturing engagement dynamics, necessitating advanced NLP and time-series forecasting techniques [6]. While Transformer-based models have shown promise in time-series analysis

[11], research on their application to non-cognitive data for predicting lecture engagement remains limited.

In this paper, we investigate the efficacy of pre-trained LLMs to predict weekly lecture-based engagement of college STEM students using their LE data. Moving beyond traditional approaches that primarily focus on cognitive attributes [8, 9], our study leverages non-cognitive data, such as students' reflections, emotions, and perceptions about lectures, supplemented by their socio-economic background information, to gain a more comprehensive understanding of engagement trends. Specifically, we address the following two research questions (RQs):

- **RQ1**: How effectively can LLMs predict STEM students' weekly lecture-based engagement using non-cognitive LE data?
- **RQ2**: To what extent does incorporating student background data alongside non-cognitive features in students' LE data enhance the accuracy of forecasting their lecture-based engagement?

For this research, we collected two years of data from 96 first-year students enrolled in an introductory programming course at a U.S. public university. The dataset includes three data modalities: background data (9 dimensions), cognitive data (41 dimensions), and non-cognitive data (28 dimensions), providing a comprehensive view of students' academic and socio-economic profiles. However, for this study, we focused on a relevant subset of the non-cognitive features and background data, excluding cognitive data, and transformed these into natural language text for processing by LLMs. Building on prior work [6], we leverage this subset to provide insights into students' academic trajectories throughout the semester. The research evaluates the effectiveness of LLMs in predicting weekly student engagement early in the semester, exploring varying data lengths and addressing the research questions related to the integration of non-cognitive and background data.

**Addressing RQ1**, we evaluate the performance of two LLM architectures: BERT (encoder-only) [12] and Llama (decoder-only) [13]. Both models are fine-tuned using four-week data periods to predict student engagement in the following week. We assess their ability to accurately predict and distinguish engagement levels, examining how each architecture handles diverse engagement patterns in educational settings.

**Addressing RQ2**, we examine the impact of integrating background data with non-cognitive features on predicting lecture-based engagement. Both BERT and Llama are fine-tuned with this enriched dataset to assess improvements in predictive accuracy. This analysis explores how socio-demographic and academic background data enhance the models' forecasting ability and lead to more precise predictions.

Our main contribution is investigating whether pre-trained LLMs can forecast college STEM students' study-related behavioral patterns, such as lecture-based engagement. We develop a robust method to achieve this by refining our previous approach to LE data enrichment, expanding the dataset to increase its scale and comprehensiveness. The integration of this enriched data with background information significantly enhances the LLMs' ability to effectively predict student engagement, demonstrating their potential in educational forecasting.

## Method

To assess whether pre-trained LLMs can accurately forecast STEM students' lecture-based engagement over the semester, we frame the forecasting task as a natural language generation problem. Although our non-cognitive dataset consists of 28 dimensions, we selectively focus on two key features—students' emotional responses before and after the lecture—based on domain knowledge and the specific needs of the forecasting task. The LLMs are fine-tuned using our processed dataset, which represents complete and meaningful sentences reflecting students' non-cognitive behaviors over time. The fine-tuning process enables the models to generate engagement predictions, classifying it as either "High" or "Low".

To handle missing values in our LE dataset, we replace absent responses with the placeholder "Skipped the question", a method introduced in our prior work [14]. This ensures missing data is processed contextually for LLMs, enhancing model performance, as shown in [15].

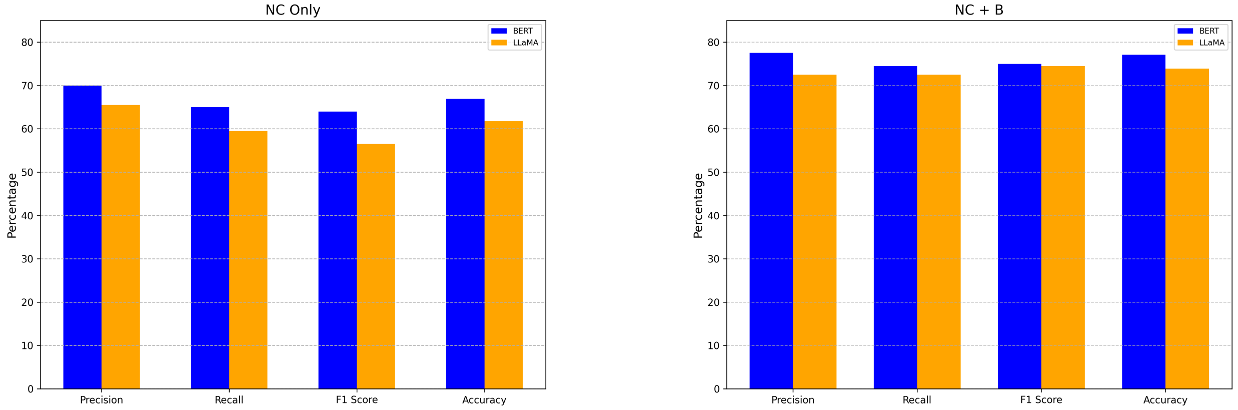
## Experiments

To address the research questions outlined in the Introduction, we designed a series of experiments to evaluate the predictive capabilities of pre-trained LLMs, specifically, the BERT-base model (encoder-only) and the Llama 3.1 8B model (decoder-only), in forecasting students' lecture-based engagement. Our objective was to assess how effectively these models leverage non-cognitive and background data to enhance prediction accuracy.

Our experiments were structured around two key feature sets: (1) non-cognitive features, including students' sentiments before and after the lecture, and (2) a combined set of non-cognitive and background data. Initially, both models were fine-tuned using only non-cognitive features to establish a baseline for prediction accuracy. We then incrementally incorporated background data to evaluate its effect on model performance. This stepwise approach allowed us to assess how the inclusion of background information influences the models' ability to distinguish between "High" and "Low" engagement levels. The results provide a nuanced understanding of the predictive strengths of encoder-only and decoder-only architectures in leveraging multimodal experiential data for engagement forecasting.

## Results

**[RQ1]:** ***How effectively can LLMs utilize non-cognitive LE data to predict college STEM students' weekly lecture-based engagement?*** We evaluated the predictive performance of two pre-trained LLMs, BERT-base (encoder-only) and Llama 3.1 8B (decoder-only), using only non-cognitive (NC) data. As shown in Figure 1 (a), BERT consistently outperformed Llama across all evaluation metrics, achieving higher precision (70.00), recall (65.00), F1-score (64.00), and accuracy (66.88). These results indicate that BERT's encoder-based architecture is more effective in capturing non-cognitive engagement patterns, likely due to its token-wise bidirectional processing, which enables richer contextual representation of input features.



(a) Performance of BERT-base and Llama 3.1 8B on Non-Cognitive Data.

(b) Performance of BERT-base and Llama 3.1 8B on Non-Cognitive + Background Data.

Figure 1: Comparison of performance on Non-Cognitive Data (left) and Non-Cognitive + Background Data (right) using BERT-base and Llama 3.1 8B.

**[RQ2]:** ***To what extent does incorporating student background data alongside non-cognitive features in students' LE data enhance the accuracy of forecasting their lecture-based engagement?*** To address RQ2, we evaluated the impact of integrating background data (NC+B) on the predictive accuracy of BERT-base and Llama 3.1 8B. As shown in Figure 1 (b), both models improved with the addition of background features. BERT achieved marked gains across all metrics: precision (77.50), recall (74.50), F1-score (75.00), and accuracy (77.07). Llama also benefited, with precision (72.50), recall (72.50), F1-score (74.50), and accuracy (73.88). These results highlight the value of background data in refining engagement predictions, with BERT maintaining a consistent performance advantage over Llama.

## Conclusion

This research demonstrates the effectiveness of pre-trained LLMs in forecasting study-related behavioral patterns, specifically lecture-based engagement, in college STEM education using LE data. By integrating non-cognitive and background features, and addressing missing values with tailored descriptors, we significantly improved predictive performance. BERT consistently outperformed Llama, particularly with non-cognitive data alone, while both models benefited from the inclusion of background information. These findings underscore the importance of contextual enrichment in forecasting student engagement. Future work can further enhance model performance by expanding the dataset, incorporating additional modalities, and refining engagement prediction methods.

## Acknowledgments

This research was supported by grants from the U.S. National Science Foundation (NSF DUE 2142558), the U.S. National Institutes of Health (NIH NIGMS P20GM130461 and NIH NIAAA R21AA029231), and the Rural Drug Addiction Research Center at the University of Nebraska-Lincoln.

## References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Dec. 2017. arXiv:1706.03762 [cs].

[2] A. Roberts, C. Raffel, and N. Shazeer, “How much knowledge can you pack into the parameters of a language model?,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 5418–5426, Association for Computational Linguistics, Nov. 2020.

[3] A. Chowdhery, S. Narang, J. Devlin, and et al., “PaLM: Scaling Language Modeling with Pathways,” oct 2022. arXiv:2204.02311 [cs].

[4] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, “Dissociating language and thought in large language models,” *Trends in Cognitive Sciences*, vol. 28, no. 6, pp. 517–540, 2024.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, pp. 140:5485–140:5551, Jan. 2020.

[6] A. Hayat, B. Khan, and M. Hasan, “Improving transfer learning for early forecasting of academic performance by contextualizing language models,” in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, (Mexico City, Mexico), pp. 137–148, Association for Computational Linguistics, June 2024.

[7] A. Hayat, S. W. Akil, H. Martinez, B. Khan, and M. R. Hasan, “Enhancing zero-shot learning of large language models for early forecasting of stem performance,” in *2024 ASEE Annual Conference & Exposition*, (Portland, Oregon), ASEE Conferences, June 2024.

[8] M. Hasan and M. Aly, “Get more from less: A hybrid machine learning framework for improving early predictions in stem education,” in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 826–831, 2019.

[9] D. Findley-Van Nostrand and R. S. Pollenz, “Evaluating psychosocial mechanisms underlying stem persistence in undergraduates: Evidence of impact from a six-day pre–college engagement stem academy program,” *CBE—Life Sciences Education*, vol. 16, no. 2, p. ar36, 2017. PMID: 28572178.

[10] M. Andrews, G. Vigliocco, and D. Vinson, “Integrating experiential and distributional data to learn semantic representations,” *Psychological Review*, vol. 116, no. 3, pp. 463–498, 2009.

[11] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, “One fits all:power general time series analysis by pretrained lm,” 2023.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.

[13] A. Grattafiori, A. Dubey, and A. J. et al., “The llama 3 herd of models,” 2024.

[14] A. Hayat, B. Khan, and M. R. Hasan, “Leveraging language models for analyzing longitudinal experiential data in education,” in *Proceedings of the 23rd International Conference on Machine Learning and Applications (ICMLA)*, (Miami, Florida), pp. 560–566, 2024.

[15] A. Hayat and M. R. Hasan, “A context-aware approach for enhancing data imputation with pre-trained language models,” in *Proceedings of the 31st International Conference on Computational Linguistics* (O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, eds.), (Abu Dhabi, UAE), pp. 5668–5685, Association for Computational Linguistics, Jan. 2025.