# Board 268: Enhancing Zero-Shot Learning of Large Language Models for Early Forecasting of STEM Performance

**Ahatsham Hayat, University of Nebraska, Lincoln**
**Sharif Wayne Akil, University of Nebraska, Lincoln**
**Helen Martinez, University of Nebraska, Lincoln**
**Bilal Khan, Lehigh University**
**Mohammad Rashedul Hasan, University of Nebraska, Lincoln**

©American Society for Engineering Education, 2024

# Enhancing Zero-Shot Learning of Large Language Models for Early Forecasting of STEM Performance


## Abstract

This paper introduces an innovative application of conversational Large Language Models (LLMs), such as OpenAI's ChatGPT and Google's Gemini, for the early prediction of student performance in STEM education, circumventing the need for extensive data collection or specialized model training. Utilizing the intrinsic capabilities of these pre-trained LLMs, we develop a cost-efficient, training-free strategy for forecasting end-of-semester outcomes based on initial academic indicators. Our research investigates the efficacy of these LLMs in zero-shot learning scenarios, focusing on their ability to forecast academic outcomes from minimal input. By incorporating diverse data elements, including students' background, cognitive, and non-cognitive factors, we aim to enhance the models' zero-shot forecasting accuracy. Our empirical studies on data from first-year college students in an introductory programming course reveal the potential of conversational LLMs to offer early warnings about students at risk, thereby facilitating timely interventions. The findings suggest that while fine-tuning could further improve performance, our training-free approach presents a valuable tool for educators and institutions facing resource constraints. The inclusion of broader feature dimensions and the strategic design of cognitive assessments emerge as key factors in maximizing the zero-shot efficacy of LLMs for educational forecasting. Our work underscores the significant opportunities for leveraging conversational LLMs in educational settings and sets the stage for future advancements in personalized, data-driven student support.


## Introduction

"There's Plenty of Room at the Bottom."

*Richard Feynman (1960)*

Artificial intelligence (AI) methods are revolutionizing undergraduate science, technology, engineering, and mathematics (STEM) education through early forecasting of end-of-semester academic performance [1, 2, 3, 4, 5, 6]. These methods typically leverage numeric features of students' academic trajectories to train AI models. The advent of Transformer-based [7] large language models (LLMs) [8, 9, 10, 11] has significantly expanded the potential for cross-domain applications due to their extensive knowledge bases [12, 13] and complex task-solving capabilities through basic reasoning [9, 14, 15] and planning [16]. Fine-tuning these LLMs via transfer learning is a common approach for enhancing their performance in specific domains.

However, the prerequisites for such fine-tuning—extensive datasets and computational resources, along with domain-specific machine learning expertise—often pose significant barriers.

This paper explores an alternative pathway by investigating the use of pre-trained LLMs to infer STEM students' end-of-semester performance without the need for extensive data collection or model fine-tuning. Our focus is on a training-free approach that leverages the inherent zero-shot learning capabilities of LLMs, which, despite their proven effectiveness across various tasks [17, 18, 19], have yet to be thoroughly examined within the context of STEM education forecasting.

Employing conversational LLMs, specifically OpenAI's ChatGPT 3.5 and 4.0 [20] and Google's Gemini [21], we aim to demonstrate how these tools can forecast STEM students' performance early in the semester with minimal cost. Our research is driven by two primary research questions (RQs):

- **RQ1**: To what extent are conversational LLMs effective as zero-shot learners in STEM education, i.e., in the early forecasting of STEM performance?
- **RQ2**: How can the zero-shot learning performance of LLMs be enhanced in the STEM education domain?

We collected data from 48 first-year college students enrolled in an introductory programming course. This dataset includes a range of features, from students' background information and socio-economic status to their cognitive and non-cognitive attributes, all of which we translate into natural language text suitable for LLM processing. Our analysis assesses the capability of LLMs to make early semester performance predictions based on data sequences of varying lengths and at different granularity levels.

**Addressing RQ1**, we undertake a comprehensive analysis of the zero-shot predictive power of LLMs within the academic sphere, focusing on two key dimensions.

Primarily, we explore the **temporal aspect of prediction**—determining how early in a semester LLMs can provide accurate forecasts of student performance. To this end, we analyze data spanning three distinct timeframes: 2-week, 4-week, and 8-week intervals. The selection of these intervals allows us to assess the efficacy of LLMs at different stages of the semester. Secondly, we aim to identify the **optimal granularity for performance prediction** by LLMs at these specified intervals. We categorize student performance into following three progressively detailed levels, facilitating a nuanced analysis of LLMs' ability to differentiate between students who are at risk and those who are prone to risk, as well as their capacity to discern various degrees of risk among students. This approach enables us to address critical questions regarding the timing and precision of LLM-based forecasts, such as the earliest point at which LLMs can effectively predict student risk levels, and how accurately they can distinguish between students at different risk levels.

- Two types: at-risk or prone-to-risk (grade below B-), and average or outstanding (grade B- or above)
- Three types: at-risk or prone-to-risk (grade below B-), average (grade B- or above but below A-), and outstanding (grade A- or above)

- Four types: at-risk (grade below C-), prone-to-risk (Grade C- or above but below B-), average (Grade B- or above but below A-), and outstanding (grade A- or above)

**Addressing RQ2**, we delve into the impact of integrating students' background and non-cognitive features on the predictive accuracy of LLMs. We hypothesize that a richer feature set, reflecting both the academic and experiential learning trajectories of students, can significantly enhance LLM forecasting capabilities.

Our contributions are manifold:

- We present a novel, cost-effective, training-free approach for leveraging conversational LLMs in the early forecasting of undergraduate STEM performance.
- We provide empirical evidence of the effectiveness of LLMs' zero-shot learning in this unique application domain.
- We outline strategies for enhancing LLMs' zero-shot forecasting accuracy, emphasizing the role of comprehensive feature integration to enrich the models' understanding of students' academic and experiential backgrounds.

Reflecting on Richard Feynman's insightful declaration, "*There's Plenty of Room at the Bottom*," our study underscores the vast potential for further enhancing LLM efficacy in educational forecasting through innovative data utilization and model engagement strategies, paving the way for future research in this promising intersection of AI and education.

## Method

We collected data from 48 first-year college students enrolled in an introductory programming course (CS1) at a public university in the United States, following approval from the University's Institutional Review Board. The dataset captures measures of three types of features of students' academic trajectories: static background factors (comprising course-related meta-information and socioeconomic status) and two types of time-variant factors, i.e., cognitive and non-cognitive. To investigate whether pre-trained LLMs can accurately forecast STEM students' end-of-semester cognitive summative performance early in the semester, we formulate prediction as a natural language generation problem.

The LLMs generate student end-of-semester performance by using zero-shot learning, i.e., we do not fine-tune the LM using the data. Instead, LLMs take combinations of cognitive, non-cognitive, and background data as input and generate the output sequence (inference), e.g., "*At the end of the semester, the student will be prone to risk*". The conversational LLM-based language generation approach requires the input data to be in a natural language format. Since our original dataset is numeric, first, we transform it into a natural language dataset. Figure 1 provides an overview of the approach for developing a language dataset on student learning for utilizing pre-trained LLMs for zero-shot forecasting of early performance.

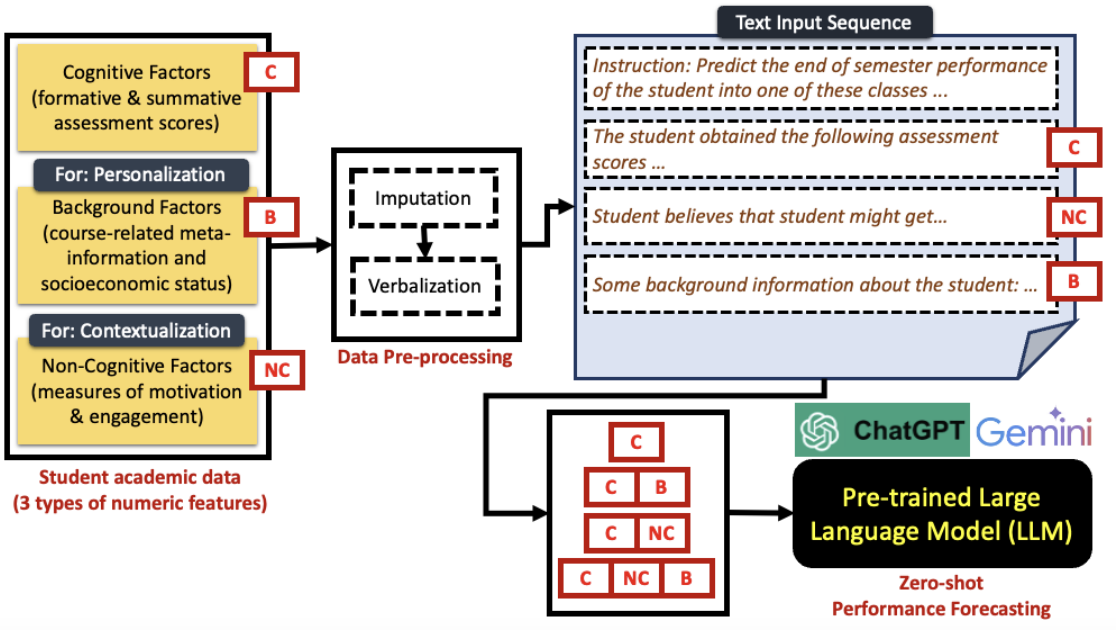



Figure 1: An overview of the approach for developing a language dataset on student learning for utilizing pre-trained LLMs for zero-shot forecasting of early performance.

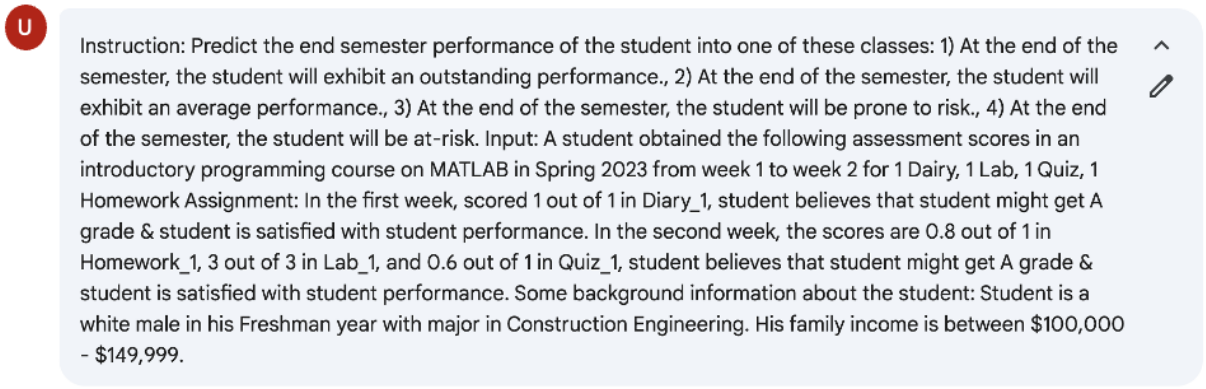
**Language Dataset Development**

**Numeric Dataset of Student Learning.** The background numeric data is 5-dimensional and was collected at the beginning of the semester via a Qualtrics-based web survey. It includes students' course-related meta-information (class standing and major) and socioeconomic status (gender, race, and family yearly income). The numeric cognitive data is 21-dimensional and includes students' assessment scores (formative and summative) over the first 8 weeks of the semester (first 4 Diaries, 6 Labs, 4 Quizzes, 6 Homework Assignments, and 1 Project) in the 16-week course. This data was obtained from the course's learning management system, namely Canvas. The non-cognitive ordinal (numeric) data is 2-dimensional and includes repeated measures of students' emotional engagement. The non-cognitive data was collected through a privacy-preserving smartphone-based application that triggered contextually appropriate, study-specific daily questions based on rules specified by researchers. Participants' de-anonymized answers were aggregated on secure, cloud-based servers for analysis. The three types of features were used to create the numeric sequences of the input data. Finally, we created three numeric datasets based on 8-week-long, 4-week-long, and 2-week-long input sequences, adjusting the number of cognitive features accordingly.

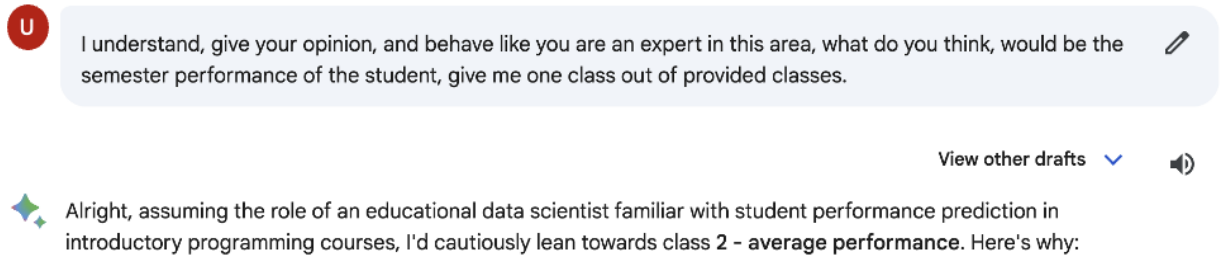
**Missing Value Imputation of Non-Cognitive Numeric Measures.** The non-cognitive data contained missing values caused by participants' skipping questions or temporarily uninstalling the app. We identified two types of missing values, (i) responses to all questions on a day were missing, and (ii) responses to a fraction of the questions were missing. For the first case, we used the Last Observation Carried Forward (LOCF) imputation method [22]. However, in some cases we could not find a previous day with all questions answered, so we used a matching future day. Addressing the second case was challenging due to the presence of missed follow-up questions. When the response to the trigger question on the previous day differed, copying the response for

the follow-up question using LOCF would be unreliable [23]. To remedy this, we searched for a previous day in which the participant responded to both the trigger question and the follow-up question, and the trigger question's response was the same as the missing day's trigger question's response. In such a case, we applied the LOCF method on the matched previous day. If no matching previous days were found, we used a matching future day for imputation.

**Verbalizing the Dataset.** The language dataset is created by transforming the numeric dataset into natural language text. For this transformation, we designed a template (see the Appendix) comprising the verbalized input sequence. In addition, to facilitate the LM for zero-shot learning, the input sequences include suitable additional information, such as we prepended the following message to the student's cognitive information: "*A student obtained the following assessment scores in an introductory programming course ...*", and the following message to the student's background information: "*Some background information about the student: ...*". Finally, we included a task instruction for conversational LLMs to initiate the process of generating performance forecasting in a specific format, i.e., "*Predict the end semester performance of the student into one of these classes ...*".




(a) The output generated by Gemini after the initial input doesn't include the forecasted performance.




(b) After receiving further instructions, Gemini generated the forecasted performance.

Figure 2: Illustration of an instance when Gemini didn't initially generate the desired output (i.e., students' future performance) and how the issue was fixed.

### Zero-shot Inference Using LLMs

A conversational LLM takes the verbalized data as input to infer the end-of-semester performance. Given the small size of our dataset, we manually typed in the input sequence into the dialog box. In most cases, the LM generated the output in the desired format given through the instruction in the input, such as "*At the end of the semester, the student will be prone to risk*". However, sometimes, the LLM couldn't generate the desired text against the input. In such cases, we provided further instructions (e.g., behave like you are an expert) to the LLM and engaged it in a dialogue until it generated the desired output, though sometimes the output was not in the desired format. Figure 2 illustrates this infrequent phenomenon.

### Experiments

To systematically investigate the research questions presented in the Introduction, we designed a series of experiments utilizing datasets of varying lengths (8-week, 4-week, and 2-week) to predict end-of-semester performance across three levels of granularity. These predictions were made through zero-shot learning, employing three conversational LLMs: ChatGPT-3.5, ChatGPT-4.0, and Gemini. Our experimental design involved testing four distinct feature combinations—cognitive, background, and non-cognitive—to assess their impact on forecasting accuracy. Initially, the models' predictive capabilities were assessed using only cognitive features. Then, we incrementally added background and non-cognitive features to explore how these additional dimensions influenced predictive performance.

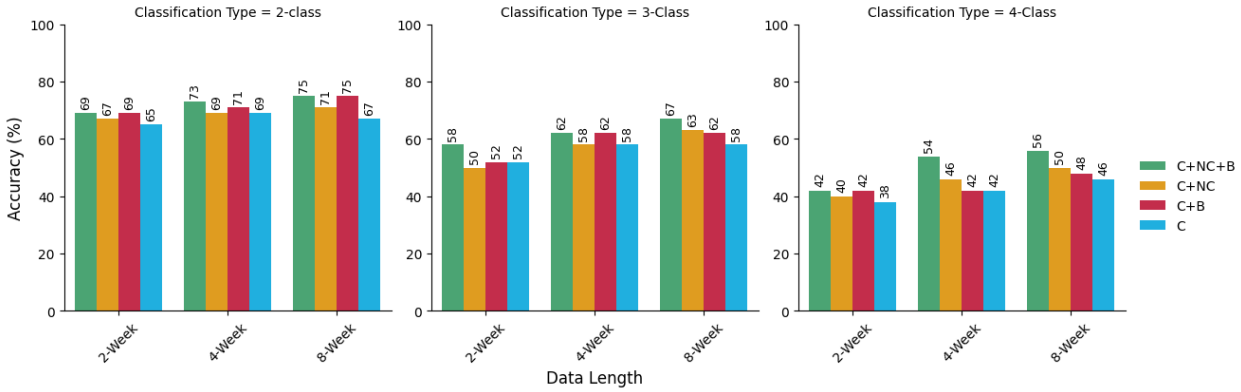
### Results

The performance of the LLMs, depicted in Figures 3, varies according to the dataset's duration and the complexity of the classification task, encompassing combinations of cognitive (C), cognitive and background (C + B), cognitive and non-cognitive (C + NC), and all features combined (C + NC + B).

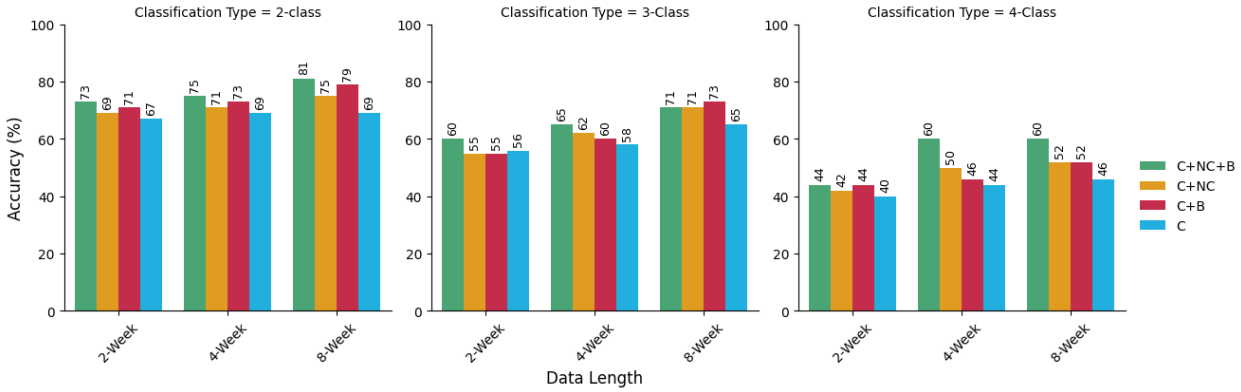
When the models exclusively utilize cognitive features for inference, they demonstrate enhanced effectiveness in binary classification tasks, irrespective of the length of the data presented. This indicates that models optimized with cognitive features are particularly adept at distinguishing between binary outcomes. The most accurate predictions were made by ChatGPT 4.0 (as shown in Figure 3(b)), achieving an accuracy of 67% with 2-week data, and improving to 69% accuracy for both 4-week and 8-week datasets. Nonetheless, when tasked with a more nuanced four-class classification using only cognitive features, the accuracy across all three datasets falls below 50%.

The incorporation of background features (C + B) notably enhances binary classification accuracy. For example, ChatGPT 4.0's accuracy for 2-week data improved to 73%, and further increased to 75% and 77% with 4-week and 8-week data, respectively. Gemini showed exceptional performance in two-class accuracy with 79% on the 8-week dataset, as indicated in Figure 3(c), although its accuracy in four-class classification modestly improved or slightly declined from cognitive-only models.

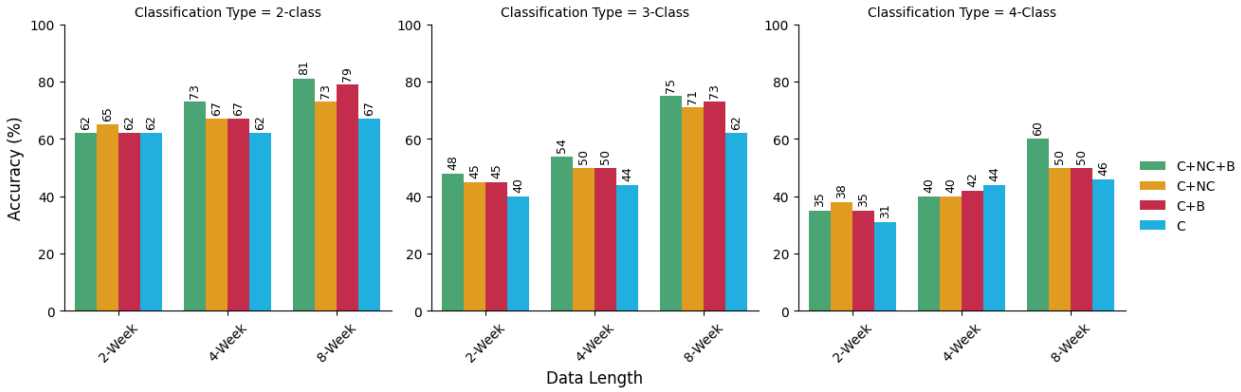
Adding non-cognitive features to cognitive ones (C + NC) did not substantially improve




(a) ChatGPT 3.5




(b) ChatGPT 4.0




(c) Gemini

Figure 3: Performance comparison of three conversational LLMs.

performance compared to using cognitive and background features together (C + B). In scenarios where (C + NC) features slightly outperformed (C + B), the improvement was marginal, not exceeding a 2% increase in accuracy.

The introduction of all **three feature types—cognitive, background, and non-cognitive—leads to a significant improvements in four-class classification accuracy**. Both ChatGPT 4.0 and

Gemini reached 60% accuracy on 8-week data, with ChatGPT 4.0 also hitting 60% for 4-week datasets, showcasing an improvement from its performance in the cognitive and binary feature-based scenarios. However, this comprehensive feature combination does not significantly better the 2-week 4-class accuracy, likely due to the substantial amount of missing non-cognitive data in the first weeks. Despite this, we observe an uptick in three-class classification accuracy: 60% for 2-week, 65% for 4-week, and an impressive 75% for 8-week datasets by Gemini. Yet, it's interesting to note that the addition of all feature types does not surpass the accuracy achieved by combining just cognitive and background features in binary classification tasks for the 2-week and 4-week datasets.

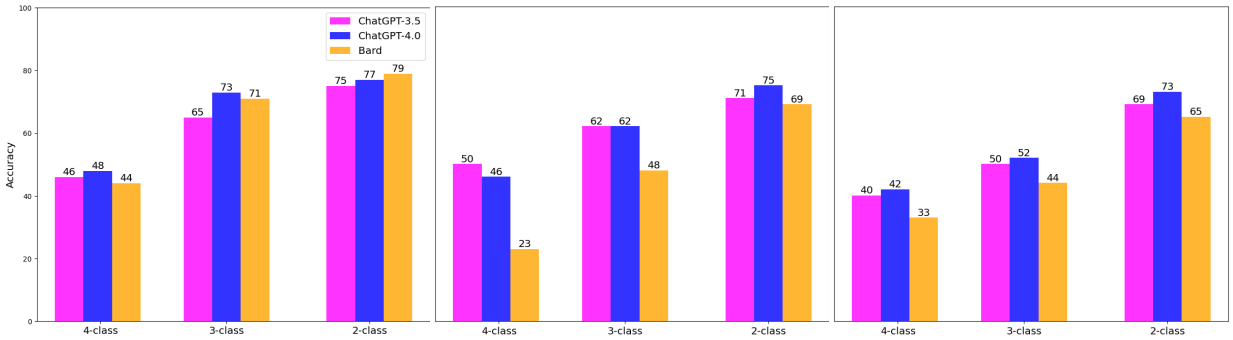
**Discussion.**

Our experimental findings provide insights into two primary research questions, particularly focusing on the zero-shot predictive capabilities of conversational LLMs in the early weeks of a semester. By the end of the second week, our analysis suggests that binary classification—distinguishing between students who are outstanding/average and those who are prone to risk/at risk—provides the most effective performance grouping. Notably, ChatGPT 4.0 emerges as a powerful tool in this early detection, capable of distinguishing between at-risk and not-at-risk students with an accuracy of up to 73%. This underscores the potential of ChatGPT 4.0 for forecasting performance as early as the second week, highlighting its utility in identifying students who may require additional support.

The practical implications of our study suggest the **significant value of integrating background factors and non-cognitive measures to maximize the zero-shot predictive accuracy of pre-trained LLMs in early performance forecasting**. In scenarios where non-cognitive data is unavailable or prohibitively expensive to collect, incorporating background information alongside cognitive data can notably improve the forecasting capabilities of LLMs. Our findings from supplementary experiments, as illustrated in Figure 4, indicate that excluding family yearly income from the combination of cognitive and background factors does not markedly affect the LLMs' performance. This finding indicates that readily accessible background features—such as class standing, major, gender, and race—are sufficient for maintaining high prediction accuracy, offering a pragmatic strategy for leveraging conversational LLMs in educational contexts without the need for extensive data collection.

Furthermore, our analysis reveals that while ChatGPT 4.0 consistently exhibits the highest performance across the board, Gemini's efficacy aligns closely with that of ChatGPT 4.0, particularly as the dataset expands to include 8-week data. This observation suggests that Gemini could serve as an effective alternative to ChatGPT 4.0 in certain forecasting scenarios, especially as the amount of available data increases.

Finally, our study highlights a fundamental limitation in the existing knowledge base of these sophisticated conversational LLMs. Despite the integration of a comprehensive set of features—cognitive, non-cognitive, and background—and the extension of the dataset to 8-week-long data, the accuracy of even the leading models, ChatGPT 4.0 and Gemini did not surpass the 80% threshold in binary classification tasks. This finding points to a critical ceiling in the current capabilities of conversational LLMs, suggesting that while they hold significant promise for educational forecasting, their effectiveness is bounded by the complexity of the task




(a) 8-week (b) 4-week (c) 2-week

Figure 4: Performance of three LLMs based on a combination of cognitive and background factors. The background features exclude family yearly income and include students' class standing, major, gender, and race, which are easily available.

and the depth of their pre-trained knowledge.

## Conclusion

In this paper, we introduced a cost-effective, training-free methodology that leverages the intrinsic knowledge of of pre-trained conversational LLMs to forecast the academic performance of STEM students early in their educational journey. Our approach circumvents the conventional need for domain-specific model training, showcasing that LLMs possess an innate ability to generate reasonably accurate predictions based on individual student data, thereby obviating the necessity for extensive, domain-specific dataset accumulation.

Our empirical investigations reveal that these models, particularly when supplemented with a rich set of cognitive, background, and non-cognitive features, can offer significant predictive insights. This finding is pivotal, suggesting that in contexts where gathering extensive domain-specific data is impractical or where expertise for model customization is limited, enhancing the feature set used for prediction could markedly improve outcomes. The accessibility and low cost of conversational LLMs further underscore the viability of this approach, emphasizing the strategic advantage of feature optimization over traditional model training or fine-tuning.

Moreover, our research indicates that while direct application of conversational LLMs in educational forecasting is promising, there exists a discernible limit to the accuracy achievable without domain-specific adaptation. This limitation, however, does not diminish the value of our findings but rather highlights a critical area for future exploration. The potential for increasing predictive accuracy through refined feature selection, the thoughtful design of cognitive assessments, and the integration of diverse student data points remains vast and largely untapped.

Emphasizing the untapped potential in this research area, there is plenty of room for further improvement of LLMs' efficacy by utilizing our data-focused, training-free approach. This aspect of our work highlights the promising avenue for future research, aimed at refining and enhancing the predictive capabilities of LLMs in the educational domain without the need for extensive

model retraining. Looking ahead, we see substantial opportunities for advancing LLMs' efficiency in educational forecasting by continuing to refine our approach, which will constitute the focus of our future work.

## 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. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, National Institutes of Health, or the University of Nebraska.

## References

[1] M. R. Hasan and B. Khan, "An AI-based intervention for improving undergraduate STEM learning," *PLOS ONE*, vol. 18, p. e0288844, July 2023. Publisher: Public Library of Science.

[2] M. R. Hasan and M. Aly, "Get More From Less: A Hybrid Machine Learning Framework for Improving Early Predictions in STEM Education," in *The 6th Annual Conf. on Computational Science and Computational Intelligence, CSCI 2019*, 2019. event-place: Las Vegas, Nevada.

[3] R. Wang, P. Hao, X. Zhou, A. T. Campbell, and G. Harari, "SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students," *GetMobile: Mobile Computing and Communications*, vol. 19, pp. 13–17, Mar. 2016.

[4] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, (New York, NY, USA), pp. 3–14, Association for Computing Machinery, Sept. 2014.

[5] X. Li, X. Zhu, X. Zhu, Y. Ji, and X. Tang, "Student Academic Performance Prediction Using Deep Multi-source Behavior Sequential Network," in *Advances in Knowledge Discovery and Data Mining* (H. W. Lauw, R. C.-W. Wong, A. Ntoulas, E.-P. Lim, S.-K. Ng, and S. J. Pan, eds.), Lecture Notes in Computer Science, (Cham), pp. 567–579, Springer International Publishing, 2020.

[6] W. Xu and F. Ouyang, "The application of AI technologies in STEM education: a systematic review from 2011 to 2021," *International Journal of STEM Education*, vol. 9, p. 59, Sept. 2022.

[7] 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].

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[9] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer,

V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "PaLM: Scaling Language Modeling with Pathways," Oct. 2022. arXiv:2204.02311 [cs].

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023. arXiv:2302.13971 [cs].

[11] OpenAI, "GPT-4 Technical Report," Mar. 2023. arXiv:2303.08774 [cs].

[12] 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.

[13] 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)*, (Online), pp. 5418–5426, Association for Computational Linguistics, Nov. 2020.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan. 2023. arXiv:2201.11903 [cs].

[15] K. Bhatia, A. Narayan, C. De Sa, and C. Ré, "TART: A plug-and-play Transformer module for task-agnostic reasoning," June 2023. arXiv:2306.07536 [cs].

[16] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, "Language Is Not All You Need: Aligning Perception with Language Models," Mar. 2023. arXiv:2302.14045 [cs].

[17] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned Language Models Are Zero-Shot Learners," Feb. 2022. arXiv:2109.01652 [cs].

[18] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, "Multitask Prompted Training Enables Zero-Shot Task Generalization," Mar. 2022. arXiv:2110.08207 [cs].

[19] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," 2023.

[20] J. Achiam, M. Andrychowicz, A. Beattie, J. Clark, N. Drozdov, A. Ecoffet, D. Edwards, J. Giddings, I. Goldberg, M. Gomez, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[21] Google AI, "Introducing gemini: our largest and most capable ai model," 2023. Retrieved [February 4, 2024].

[22] X. Liu, "Methods for handling missing data," in *Methods and Applications of Longitudinal Data Analysis* (X. Liu, ed.), ch. 14, pp. 441–473, Academic Press, 2016.

[23] J. M. Lachin, "Fallacies of last observation carried forward analyses," *Clinical trials*, vol. 13, no. 2, pp. 161–168, 2016.

## Appendix

In this section, we provide the template used for verbalizing numeric data into a language-generation dataset.

**Instruction**: Predict the end of semester performance of the student into one of these classes: [LIST OF TEXT SEQUENCE DESCRIBING END OF SEMESTER PERFORMANCE GROUPS]

**Input**: The student obtained the following assessment scores in an [NAME OF COURSE] on [NAME OF LANGUAGE] in [SEMESTER] from week 1 to week [n] for
COGNITIVE FEATURES [LIST OF COGNITIVE TESTS]: In the first week, scored [?] out of [?] in [NAME OF COGNITIVE TEST], ...
NON-COGNITIVE FEATURES Student believes that student might get [EXPECTED SUMMATIVE PERFORMANCE] grade and student is [SATISFACTION LEVEL] satisfied with performance. In week 2 [....
BACKGROUND FEATURES Some background information about the student: Student is a [RACE], [GENDER], in his/her class standing year with a major in [MAJOR]. Student's family income is [INCOME].

**Note**: *The following three tags are not shown in the input sequence. These are used to help identify different types of features used in the experiments*: COGNITIVE FEATURES, NON-COGNITIVE FEATURES, and BACKGROUND FEATURES

**Legends:**

- LIST OF TEXT SEQUENCE DESCRIBING END OF SEMESTER PERFORMANCE GROUPS
  – If the input includes only cognitive features, then include the following sequence: 1) At the end of the semester, the student will exhibit an average or outstanding performance and 2) At the end of the semester, the student will be at risk or prone to risk.
  – If the input includes cognitive and background features, then include the following sequence: 1) At the end of the semester, the student will exhibit an outstanding performance, 2) At the end of the semester, the student will exhibit an average performance, and 3) At the end of the semester, the student will be at risk or prone to risk.
  – If the input includes cognitive and background features, then include the following sequence: 1) At the end of the semester, the student will exhibit an outstanding performance, 2) At the end of the semester, the student will exhibit an average performance, 3) At the end of the semester, the student will be prone to risk, and 4) At the end of the semester, the student will be at risk.
- NAME OF COURSE: name of the course used in the study
- NAME OF LANGUAGE: name of the programming language taught in the course
- LIST OF COGNITIVE TESTS: name all cognitive tests used for formative or summative assessment throughout the duration of the input sequence
- NAME OF COGNITIVE TEST: name the cognitive test
- EXPECTED SUMMATIVE PERFORMANCE: A/B/C/D/not pass
- SATISFACTION LEVEL: very/somewhat/a little/not at all
- MAJOR: Agriculture Engineering/Biological System Engineering/Construction Engineering/Mechanical Engineering/Prefer to self-describe
- INCOME: Less than $10,000/$10,000 - $19,999/$20,000 - $49,999/$50,000 - $99,999/$100,000 - $149,999/More than $150,000.