

# Leveraging Language Models for Analyzing Longitudinal Experiential Data in Education

Ahatsham Hayat

Electrical and Computer Engineering  
University of Nebraska-Lincoln  
aahatsham2@huskers.unl.edu

Bilal Khan

Computer Science and Engineering  
Lehigh University  
bik221@lehigh.edu

Mohammad Rashedul Hasan

Electrical and Computer Engineering  
University of Nebraska-Lincoln  
hasan@unl.edu

**Abstract**—We propose a novel approach to leveraging pre-trained language models (LMs) for early forecasting of academic trajectories in STEM students using high-dimensional longitudinal experiential data. This data, which captures students' study-related activities, behaviors, and psychological states, offers valuable insights for forecasting-based interventions. Key challenges in handling such data include high rates of missing values, limited dataset size due to costly data collection, and complex temporal variability across modalities. Our approach addresses these issues through a comprehensive data enrichment process, integrating strategies for managing missing values, augmenting data, and embedding task-specific instructions and contextual cues to enhance the models' capacity for learning temporal patterns. Through extensive experiments on a curated student learning dataset, we evaluate both encoder-decoder and decoder-only LMs. While our findings show that LMs effectively integrate data across modalities and exhibit resilience to missing data, they primarily rely on high-level statistical patterns rather than demonstrating a deeper understanding of temporal dynamics. Furthermore, their ability to interpret explicit temporal information remains limited. This work advances educational data science by highlighting both the potential and limitations of LMs in modeling student trajectories for early intervention based on longitudinal experiential data.

**Index Terms**—Language Model, Time-series Data, Experiential Data, Missing Data, Data Augmentation, STEM Education

## I. INTRODUCTION

“We apprehend more than we comprehend.”

*Michel Serres, The Parasite (1982)*

Experience plays a central role in the human learning process [1]. Gaining insights into learners' experiences can significantly enhance learning outcomes [2], [3]. This is often achieved by analyzing **longitudinal experiential data**, which involves systematically aggregating real-time observations from individuals over time through methods such as self-report surveys. This data encompasses learners' perceptions and interactions within their learning activities [4]. Such rich, time-varying data on human experience provides insights into the subjective and qualitative dimensions of learning, including emotions, perceptions, opinions, and values related to their educational engagements. Additionally, it captures extralinguistic aspects of a learner's journey, uncovering their strengths, weaknesses, opportunities, and challenges. This, in turn, offers pathways for enhancing educational experiences. For example, understanding students' academic experiences through this data can inform just-in-time intervention strategies,

potentially predicting cognitive performance well before the end of the semester [5]–[8].

Leveraging longitudinal experiential data to forecast academic performance with artificial intelligence (AI), particularly through deep learning (DL) techniques, presents notable challenges. Firstly, since experiential data primarily consists of qualitative text, advanced DL-based natural language processing (NLP) techniques are required to effectively interpret and utilize this non-numeric information. Secondly, the inherent temporal dynamics within longitudinal experiential data, characterized by repeated measurements of experiential attributes, transform it into a complex time-series dataset. This complexity necessitates the development of innovative DL-based approaches for time-series forecasting that are specifically designed to accommodate and learn from qualitative inputs.

Recently, Transformer-based [9] pre-trained language models (LMs) [10], [11] have revolutionized various AI domains, including time-series forecasting [12]–[18]. However, the suitability of these models for creating intervention systems based on qualitative longitudinal experiential data remains under-explored. Most existing methods are tailored to numeric, non-experiential longitudinal data, indicating a gap in the application of pre-trained LMs to the nuanced and text-rich domain of experiential data.

In this research, we explore the extent to which pre-trained LMs can effectively interpret and utilize longitudinal experiential data within the context of student learning, specifically in STEM (science, technology, engineering, and mathematics) education. Addressing the lack of suitable high-dimensional experiential datasets for such research, we have compiled a unique **78-dimensional dataset** that encompasses a holistic view of college students' academic journey over a semester.

This dataset is divided into three key components: (i) The non-cognitive component, which includes 28 dimensions of repeated qualitative measurements, captures attributes such as student motivation and engagement, offering insights into students' perceptions of their academic experiences; (ii) The cognitive component, consisting of 41 quantitative measures, encompasses students' formative and summative assessment scores; and (iii) The background factors component, featuring 9 dimensions of qualitative data, provides static information on students' academic meta-information and socioeconomic status.

Crucially, both the non-cognitive and cognitive data components are structured as **time-series**. Utilizing this comprehensive, high-dimensional dataset, our research aims to determine if pre-trained LMs can effectively decipher experiential cues for early forecasting of students' end-of-semester cognitive performance. Specifically, we examine the models' ability to learn and integrate the complex correlations between non-cognitive and cognitive data elements, and to account for their temporal variations. This exploration is pivotal in understanding how advanced LMs can adapt to the nuanced domain of educational data, which encompasses both qualitative and quantitative aspects, thereby facilitating the development of effective just-in-time interventions.

The nature of our longitudinal experiential student data presents **unique challenges**, distinctly setting it apart from the time-series numeric data typically employed by state-of-the-art time-series LM-based methods [12]. Our dataset's distinctive characteristics include: (i) A hybrid structure combining static background features with time-variant cognitive and non-cognitive elements; (ii) The inclusion of non-numeric, experiential measurements, (iii) The forecasted variable is text-based assessments of future summative performance; (iv) A significant proportion of missing values within the non-cognitive data, complicating the learning of correlations with the time-series cognitive components (discussed in Section II); (v) A lack of temporal alignment between non-cognitive and cognitive modalities, with respective cross-modality features often recorded on different days; (vi) A relatively small dataset size ( $N=48$ ), posing challenges for effective LM-based transfer learning due to the high cost of collecting comprehensive longitudinal data.

To address these challenges and harness the general knowledge and reasoning capabilities of pre-trained LMs [19]–[22], we develop a **data enrichment** method. This approach enables fine-tuning pre-trained LMs for early performance forecasting by reframing cognitive performance forecasting as a language generation problem. Our data enrichment method involves: (i) Handling of missingness in student experiential data; (ii) Augmentation of text sequence data to counter the limitations posed by the dataset's small size; (iii) Inclusion of explicit task instructions and contextual cues to guide LMs in understanding the task, recognizing temporal orders, and applying domain-specific knowledge, thereby addressing learning challenges across different data dimensions.

In our comprehensive empirical study, we evaluate two types of pre-trained LMs – decoder-only and encoder-decoder models – to systematically examine their capability in early forecasting of summative cognitive performance by utilizing experiential data. We address the following pivotal research questions (RQs):

- **[RQ1]:** To what extent can LMs accurately forecast outcomes based solely on longitudinal experiential data?
- **[RQ2]:** How effectively do LMs capture and leverage correlations across the non-cognitive, cognitive, and background modalities within academic experiential data for precise early forecasting?

- **[RQ3]:** What is the extent of LMs' ability to interpret and use temporal variations within the dataset for forecasting purposes?
- **[RQ4]:** How can we effectively address missingness in experiential datasets by leveraging pre-trained LMs?

Our key contributions include the creation of a multi-dimensional longitudinal experiential dataset focused on STEM education, the development of a novel data enrichment method tailored for pre-trained LMs, and a set of extensive empirical studies that shed light on the learning behaviors of decoder-only and encoder-decoder LMs. A significant observation from our studies is the impressive capability of LMs to assimilate experiential data and effectively discern correlations across the modalities of experiential data. However, while LMs are adept at identifying temporal patterns in time-series data, they tend to rely predominantly on surface-level statistical relationships. For instance, our findings indicate that LMs do not effectively utilize explicit temporal tags embedded in the dataset to learn time-series patterns. This limitation underscores that within the scope of longitudinal experiential data, LMs excel more at uncovering complex statistical correlations than at comprehending deeper semantic contexts, a crucial aspect for the advancement of just-in-time intervention systems in education using LMs.

## II. DATASET DEVELOPMENT & ENRICHMENT

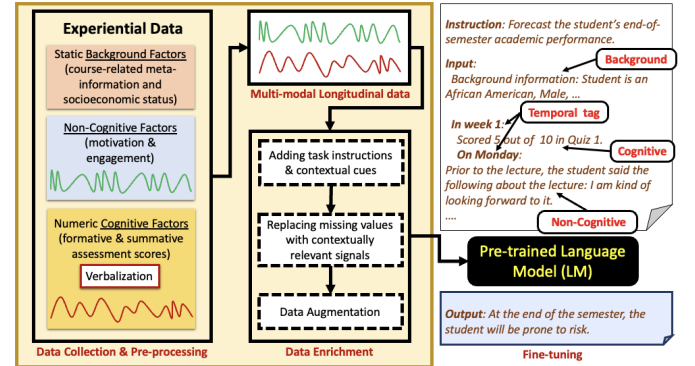


Fig. 1: Overview of the longitudinal experiential dataset development process that is amenable to adapting LMs.

To enable pre-trained LMs to gain a nuanced understanding of students' academic experiences, we compiled a comprehensive dataset that explores the interplay between experiential modalities in student learning. Figure 1 provides an overview of our dataset development process, including data collection, pre-processing, and a method for data enrichment. The transformed data is utilized to adapt pre-trained LMs for early forecasting of students' end-of-semester cognitive performance based on the first 4 weeks of data. Performance is categorized into four types, aligned with major letter grade thresholds: at-risk (grade C or below), prone-to-risk (above C but below B), average (above B but below A), and outstanding (grade A or above).

### A. Data Collection

We gathered data from 48 first-year college students enrolled in an introductory programming course at a public university in

the U.S. The dataset, comprising 78-dimensional data, captures three modalities of students’ academic experiential trajectories, as described below.

**Background Data (9-dimensional):** At the semester’s start, we collected essential 9-dimensional background data via a Qualtrics-based web survey. This data includes course-related meta-information (class standing and major) and socioeconomic factors (gender, race, international or native student status, parents’ education background, highest education level of a single parent, highest education level of another parent, family yearly income, science identity, and reflected science identity). These factors are included to examine the **impact of personalization** on LMs, hypothesizing that these attributes correlate with students’ academic trajectories and future course performance [23], serving as valuable priors for LMs to recognize individualized patterns in academic progression.

**Cognitive Data (41-dimensional):** This data includes 41-dimensional cognitive data from students’ assessment scores (formative and summative) throughout the 16-week semester, sourced from the course’s learning management system.

**Non-Cognitive Data (28-dimensional):** The 28-dimensional non-cognitive data comprises repeated measures of students’ motivation (intrinsic and extrinsic) and engagement (behavioral, emotional, and cognitive) factors across the semester. This selection aims to capture students’ evolving study-related behaviors, with research indicating a strong correlation between these non-cognitive factors and students’ learning outcomes [24], [25]. This correlation is essential for an LM to effectively capture subtle variations in academic performance that may not be discernible solely from their cognitive trajectory data. The non-cognitive data was sourced from a smartphone-based application. The privacy-preserving app triggered contextually tailored, study-specific daily questions, following rules stipulated by researchers. Participants’ anonymized responses were securely aggregated on cloud-based servers for subsequent analysis.

### B. Data Pre-processing

The cognitive data, represented numerically, and the background and non-cognitive data, expressed in natural language, required alignment. Thus, we **verbalized** the cognitive data, converting numerical scores into natural language descriptions. For instance, scores of 1/1, 3/3, and 0.8/1 in Homework 1, Lab 1, and Quiz 1, respectively, were verbalized as “*The scores are 1 out of 1 in Homework\_1, 3 out of 3 in Lab\_1, and 0.8 out of 1 in Quiz\_1.*” This verbalization enables the integration of cognitive data with the text-based background and non-cognitive data. Specifically, static background text data was prepended to the longitudinal cognitive and non-cognitive data to form the input text sequence  $X$ .

Given the large 78-dimensional feature space and the limitations of the input context window sizes of the LMs used (e.g., the encoder-decoder LM FLAN-T5 [26] used in this research can only accommodate 512 tokens), we selected a subset of features to keep the number of tokens within

512. This selection includes 5-dimensional distal background factors (class standing, major, gender, race, and family yearly income), 10-dimensional cognitive factors spanning over the first 4 weeks of the semester (first 2 Diaries, 3 Labs, 2 Quizzes, and 3 Homework Assignments), and 3-dimensional experiential non-cognitive factors (i.e., repeated measures of students’ three types of engagement factors—behavioral, emotional, and cognitive). We used responses from Monday, Thursday, and Saturday.

The output text sequence  $Y$  reflects the student’s end-of-semester final letter grade, categorized into four performance groups. Finally, the input and output data sequences were combined to create a language dataset. To assess how early in the semester LMs can accurately forecast performance, we created datasets based on 2-week, 3-week, and 4-week-long input sequences, adjusting the number of cognitive features accordingly.

### C. Data Enrichment

Our data enrichment method is designed to enhance LM adaptation, comprising task instructions, contextual cues, a strategy for handling missing values, and dataset augmentation.

**Task Instructions and Contextual Cues:** We incorporated a task instruction at the start of each input data sequence  $X$ , which reads: “*Forecast the student’s end-of-semester academic performance.*” This instruction is crucial in adapting the LM for fine-tuning based on instructional cues [27]. Additionally, **contextual messages** were incorporated for clarity, such as “*Background information:*” at the start of background information, and temporal cues like “*In week [WEEK\_NUMBER]*” and “*On [NAME\_OF\_THE\_DAY]*” for weekly and daily data, respectively. The output sequence  $Y$  is contextualized with expressions like “*At the end of the semester, the student will be [STUDENT’S\_PERFORMANCE]*”

**Replacing Missing Values with Contextually Relevant Descriptors.** Our longitudinal experiential dataset, which includes weekly responses to three non-cognitive questions, exhibited a considerable incidence of missing values. These gaps primarily arose when participants either skipped questions or temporarily uninstalled the data-collection app. In the initial week, for instance, students omitted responses to 66% of the non-cognitive questions. Moreover, more than 37% of participants skipped at least one such question over a period extending beyond two weeks. To address this issue, we eschewed traditional data imputation strategies in favor of inserting a contextually relevant descriptor text, specifically “*Skipped the question*”, wherever data was absent. This approach was congruent with the nature of how missing values manifested in our dataset. Our decision to refrain from standard imputation methods, like Last Observation Carried Forward (LOCF) [28], was informed by scenarios where entire sets of daily responses were missing, rendering methods like LOCF unsuitable.

**Augmenting the Language Dataset.** To address the unbalanced distribution in the initial dataset (out of 48 instances 24 outstanding, 12 average, 6 prone-to-risk, and 6 at risk), we

employed oversampling with random sampling techniques [29] and synonym replacement for token variation [30]. This resulted in a near-balanced distribution of performance categories, reducing potential biases in LM predictions. The augmented dataset comprises 144 samples (48 outstanding, 36 average, 30 prone-to-risk, and 30 at-risk).

### III. EXPERIMENTS

This section presents a series of experiments designed to investigate four key research questions about the learning behaviors of LMs, as detailed in Section I. Our experimental setup involved two distinct types of pre-trained LMs: the decoder-only LLaMA 2 (Large Language Model Meta AI) [31] and the encoder-decoder FLAN-T5 [26]. These selections allowed us to compare the performance between a large-scale LM and a moderately-sized LM, specifically the 7 billion-parameter (7B) LLaMA 2 and the 770 million-parameter (770M) FLAN-T5. We fine-tuned these models across three different language datasets of varying lengths: 4-week, 3-week, and 2-week durations, aiming to assess the adaptability of LMs over diverse time frames. The performance of these adapted LMs was evaluated based on their ability to generate outputs with matching keywords corresponding to predefined performance types.

**Test Datasets.** For our testing purposes, we curated datasets by sampling approximately 30% of instances from the augmented datasets, ensuring a balanced distribution across different classes. The rest, constituting 70% of the data, was employed for the fine-tuning process of the LMs.

**Experimental Setup.** For the decoder-only LM, we utilized the 7B LLaMA 2 model [31], characterized by a maximum token limit of 4,096 in its context window. We fine-tuned this model using a parameter-efficient fine-tuning (PEFT) method QLoRA [32] with the following parameter settings: `lora_r = 16`, `lora_alpha = 64`, `lora_dropout = 0.1`, `task_type = "CAUSAL_LM"`. The model's learning rate was set to  $2e-4$ , and the optimizer used was `paged_adamw_32bit`.

In the case of the encoder-decoder LM, the 770M FLAN-T5 model [26], a variation of the T5 model [19], was selected. This model has a context window capping at 512 tokens. We fine-tuned the FLAN-T5 using an AdamW optimizer [33] with a learning rate set to  $3e-4$ .

All experiments were conducted with a batch size of 4, spanning 50 epochs. This batch size was chosen in consideration of the memory limitations encountered during the fine-tuning phase. We utilized Tesla V100 (32GB RAM) and A40 (48GB RAM) GPUs for distributed training. The fine-tuning process for each model was completed in under an hour.

#### A. Results

**[RQ1]: To what extent can LMs accurately forecast outcomes based solely on longitudinal experiential data?**

To explore RQ1 and RQ2, we examined the forecasting abilities of two types of LMs (LLaMA 2 and FLAN-T5) using both the three modalities of longitudinal experiential data. Our

approach involved fine-tuning the LMs across five distinct combinations of the three data modalities and employing language datasets of varying durations (4-, 3-, and 2-week).

Our findings, illustrated in Figure 2, reveal a notable trend: models fine-tuned exclusively with non-cognitive data consistently outperformed those trained solely on cognitive-only data. This outcome underscores the LMs' remarkable ability to extract meaningful insights into students' academic progress by analyzing their experiential data. What makes this even more significant is the limited feature set used for adaptation – only three out of twenty-eight available features related to motivation and engagement were employed.

This performance was particularly striking in the case of the 2-week language datasets. Despite grappling with a 66% rate of missing values in the first week's non-cognitive data, the LMs demonstrated impressive forecasting accuracy for end-of-semester cognitive performance. The accuracy rates exceeded 70% on average, with LLaMA achieving 74.0% and FLAN-T5 reaching 71.4%. In contrast, the cognitive-only models lagged slightly behind, with LLaMA at 71.4% and FLAN-T5 at 67.4%.

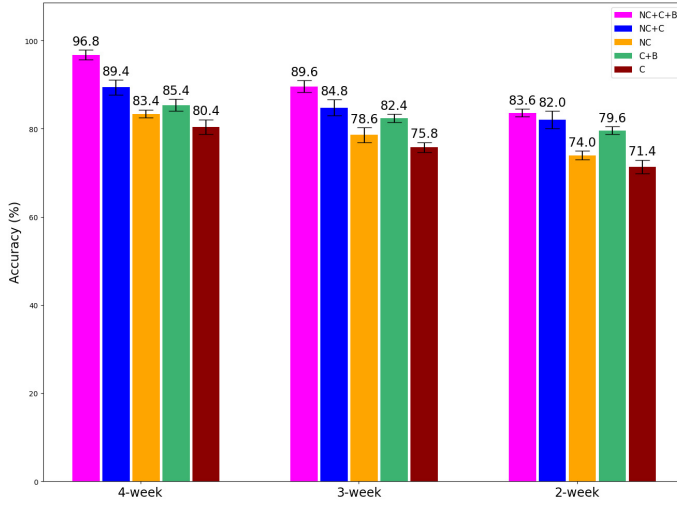
**[RQ2]: How effectively do LMs capture and leverage correlations across the non-cognitive, cognitive, and background modalities within academic experiential data for precise early forecasting?**

Our initial hypothesis posited that the key to accurate forecasting lies in the LMs' ability to learn and integrate correlations across data modalities. The results, as depicted in Figure 2, affirm this hypothesis. We observed that both LLaMA and FLAN-T5 LMs achieved their peak performance when fine-tuned with a combination of three modalities, highlighting the importance of a multi-modal approach.

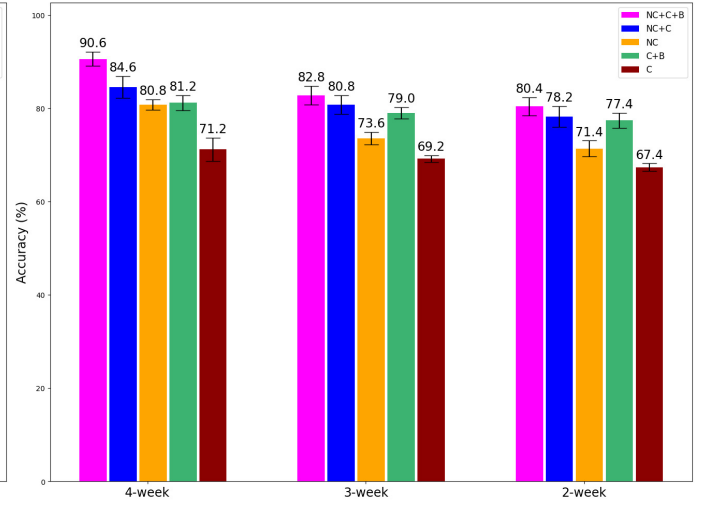
In terms of comparative performance, LLaMA consistently outshone FLAN-T5. A distinguished milestone was reached as early as week 2, with the LLaMA model achieving an impressive 83.6% accuracy in its forecasts, while FLAN-T5 followed closely with 80.4% accuracy. The peak of this performance was observed with the week-4 data, where LLaMA reached a remarkable 96.8% accuracy. This high level of accuracy achieved by LLaMA underscores the efficacy of our approach in leveraging the full spectrum of data modalities for early and precise academic performance forecasting.

**[RQ3]: What is the extent of LMs' ability to interpret and use temporal variations within the dataset for forecasting purposes?**

The impressive accuracy demonstrated by the LMs (refer to Figure 2) raises an intriguing question: Do these models genuinely grasp and utilize the temporal dynamics in the dataset, or are they merely capturing broad statistical patterns? This inquiry is crucial, considering that our dataset was enriched with contextual temporal markers to signify the chronological progression of weeks (e.g., "In week 1") and days within a week (e.g., "On Monday"). We anticipated that these temporal cues would be instrumental for the LMs in discerning and leveraging the nuanced variations in the time-series data.

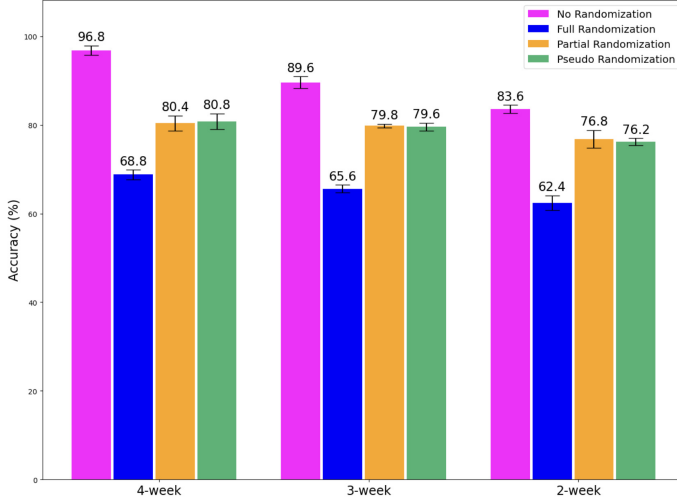


(a) LLaMA

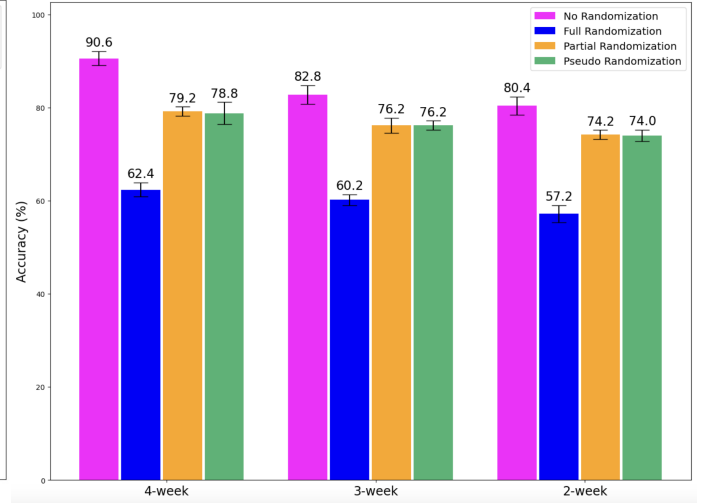


(b) FLAN-T5

Fig. 2: [RQ1 & RQ2]: Evaluation of two types of LMs that are fine-tuned with various combinations of experiential and non-experiential modalities of data using the 4-week, 3-week, and 2-week datasets. Each model is evaluated 5 times and the average and standard deviation are reported. *Legends: NC=Non-Cognitive, C=Cognitive, B=Background.*



(a) LLaMA



(b) FLAN-T5

Fig. 3: [RQ3]: Performance of the LMs (fine-tuned with both modalities) is compared with LMs fine-tuned with randomized measures of three experiential modalities.

To probe deeper into this aspect, we have crafted a trio of experiments, each utilizing a different iteration of the dataset that gradually enhances the degree of temporal randomization. Moreover, in the first two experiments, we intentionally omit the temporal markers to assess the models' ability to perceive temporal sequences without explicit cues. These experiments are designed to unravel the extent to which the LMs are sensitive to the temporal ordering of data, whether they can still perform effectively when this order is disrupted, and crucially, their proficiency in harnessing explicit temporal cues when available.

- Experiment RQ3(a) [**Full randomization**]: This experiment introduces complete randomization by shuffling both the order of weeks and days within each week.

Additionally, we omit temporal markers such as “In Week [WEEK\_COUNT]” to eliminate any explicit chronological cues.

- Experiment RQ3(b) [**Partial randomization**]: Here, we only randomize the sequence of weeks while maintaining the day-to-day order within each week. Similar to RQ3(a), we remove weekly temporal tags to obscure the original temporal sequence.
- Experiment RQ3(c) [**Pseudo randomization**]: This setup mirrors the partial randomization, except that we retain the weekly tags. Despite the shuffled order of weeks, these tags could potentially aid the models in discerning the chronological sequence, hence the term “pseudo randomization.”

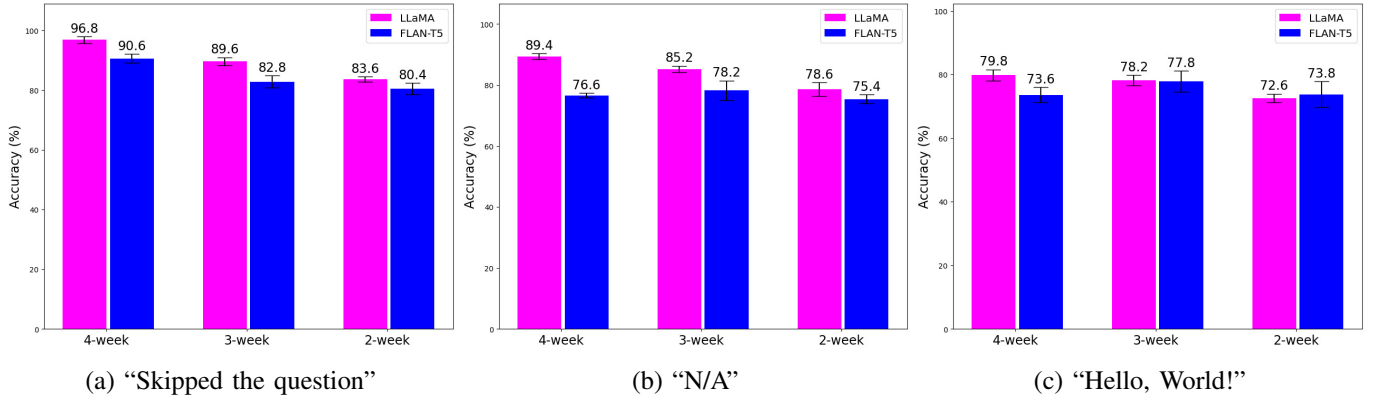


Fig. 4: [RQ4]: Investigation of the impact of the contextually relevant descriptor of missing values.

For both LLaMA and FLAN-T5 models, we evaluated their performance on these three randomized datasets against a baseline trained on the original, chronologically ordered dataset (no randomization). The results, depicted in Figure 3, are revealing. With full randomization (Experiment RQ3(a)), there’s a significant performance drop (ranging between 21% to 28% across different time frames) compared to the baseline. This decline highlights the impact of temporal disarray on the models’ forecasting abilities.

In the partial randomization scenario (Experiment RQ3(b)), the performance decrement is less severe, under 10% for both models across 2-week and 3-week datasets. However, for the 4-week dataset, the decline is more pronounced, reaching up to 16.4% for LLaMA and 11.4% for FLAN-T5. This trend suggests that retaining daily sequences within each week mitigates the negative impact of disrupted weekly sequences, though not completely.

Interestingly, in the pseudo randomization experiment (RQ3(c)), despite preserving the weekly tags, there was no significant improvement in model accuracy over the partially randomized scenario. In some cases, such as with the LLaMA model, performance even slightly declined (by 0.2% in 3 weeks and 0.6% in 2 weeks). Thus, even with its comparatively larger base of general knowledge and better reasoning ability, the 7B LLaMA does not seem to pick up the explicitly encoded temporal signals. This outcome suggests that **while the models are capable of learning statistical patterns from time-varying measures, they may struggle to fully comprehend and utilize explicit temporal cues encoded in the data.**

**[RQ4]: How can we effectively address missingness in experiential datasets by leveraging pre-trained LMs?**

A significant challenge in our dataset is the presence of numerous missing values in its experiential dimension. We explore how the general knowledge embedded in LMs can be harnessed to address this issue. Prior research has demonstrated that LMs are adept at managing missing values, often by substituting them with a generic descriptor like “N/A” [14]. Our focus here is on evaluating whether contextually nuanced descriptors for missing data enhance the performance of LMs in comparison to standard descriptors, and to what extent incorrect descriptors influence model accuracy.

Three experiments were designed to investigate this:

- Experiment RQ4(a) [**Replacement with “Skipped the question”**]: This experiment (Figure 4(a)) involved substituting missing values with the phrase “Skipped the question”, a contextually relevant descriptor.
- Experiment RQ4(b) [**Replacement with “N/A”**]: In this setup (Figure 4(b)), missing values were replaced with the more generic but still contextually correct “N/A”.
- Experiment RQ4(c) [**Replacement with “Hello, World!”**]: This experiment (Figure 4(c)) used “Hello, World!” as an intentionally contextually incorrect descriptors for missing values.

The results reveal some intriguing patterns. The first experiment (RQ4(a)) using “Skipped the question” led to the highest performance in both LLaMA and FLAN-T5 models, underscoring the value of contextually rich descriptors. Surprisingly, even though “N/A” is contextually correct, its use resulted in a marked performance drop (RQ4(b)). For instance, accuracy for the 4-week LLaMA model decreased by 7.4%, and for the FLAN-T5 model, the decline was even more significant at 14%. This suggests a sensitivity of these models to the specific wording used in missing value descriptors.

Furthermore, in RQ4(b) and RQ4(c), despite using an incorrect descriptor, the performance decrease was minor: less than 3% for FLAN-T5 and less than 10% for LLaMA. This suggests that LMs maintain robustness even with missing data, regardless of the descriptor’s contextual accuracy.

#### IV. CONCLUSION

This research explores the capabilities of pre-trained LMs for early forecasting of academic trajectories in STEM students using longitudinal experiential data. Leveraging a novel dataset that encompasses non-cognitive, cognitive, and background factors, we fine-tuned LMs through an innovative data enrichment process that addressed missing values, augmented textual sequences, and incorporated task-specific instructions and contextual indicators. Our findings reveal that while LMs effectively integrate multiple data modalities and demonstrate robustness in handling incomplete data, they primarily rely on high-level statistical patterns, lacking deeper semantic understanding of temporal dynamics. Additionally, their ability



to interpret explicit temporal information remains limited. Moving forward, expanding the dataset will be critical to improving fine-tuning and gaining deeper insights, particularly into non-cognitive features.

#### 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] D. A. Kolb, *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [2] M. Palmer, M. Larkin, R. de Visser, and G. Fadden, "Developing an interpretative phenomenological approach to focus group data," *Qualitative Research in Psychology*, vol. 7, no. 2, pp. 99–121, 2010. [Online]. Available: <https://doi.org/10.1080/14780880802513194>
- [3] S. Petrina, "Methods of analysis experiential analysis," 2018. [Online]. Available: <https://blogs.ubc.ca/educ500/files/2019/02/Experiential-Analysis.pdf>
- [4] 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. [Online]. Available: <https://doi.org/10.1037/a0016261>
- [5] 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, no. 4, pp. 13–17, Mar. 2016. [Online]. Available: <https://doi.org/10.1145/2904337.2904343>
- [6] 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*, ser. UbiComp '14. New York, NY, USA: Association for Computing Machinery, Sep. 2014, pp. 3–14. [Online]. Available: <http://doi.org/10.1145/2632048.2632054>
- [7] 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*, ser. Lecture Notes in Computer Science, H. W. Lauw, R. C.-W. Wong, A. Ntoulas, E.-P. Lim, S.-K. Ng, and S. J. Pan, Eds. Cham: Springer International Publishing, 2020, pp. 567–579.
- [8] 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, no. 1, p. 59, Sep. 2022. [Online]. Available: <https://doi.org/10.1186/s40594-022-00377-5>
- [9] 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]. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [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]. [Online]. Available: <http://arxiv.org/abs/2302.13971>
- [11] OpenAI, "GPT-4 Technical Report," Mar. 2023, arXiv:2303.08774 [cs]. [Online]. Available: <http://arxiv.org/abs/2303.08774>
- [12] M. Jin, Q. Wen, Y. Liang, C. Zhang, S. Xue, X. Wang, J. Zhang, Y. Wang, H. Chen, X. Li, S. Pan, V. S. Tseng, Y. Zheng, L. Chen, and H. Xiong, "Large models for time series and spatio-temporal data: A survey and outlook," 2023.
- [13] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, "One fits all: power general time series analysis by pretrained lm," 2023.
- [14] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," 2023.
- [15] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, "Time-llm: Time series forecasting by reprogramming large language models," 2023.
- [16] C. Chang, W.-C. Peng, and T.-F. Chen, "Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms," 2023.
- [17] D. Cao, F. Jia, S. O. Arik, T. Pfister, Y. Zheng, W. Ye, and Y. Liu, "Tempo: Prompt-based generative pre-trained transformer for time series forecasting," 2023.
- [18] C. Sun, Y. Li, H. Li, and S. Hong, "Test: Text prototype aligned embedding to activate llm's ability for time series," 2023.
- [19] 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, no. 1, pp. 140:5485–140:5551, Jan. 2020.
- [20] 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: Association for Computational Linguistics, Nov. 2020, pp. 5418–5426. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.437>
- [21] 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]. [Online]. Available: <http://arxiv.org/abs/2201.11903>
- [22] K. Bhatia, A. Narayan, C. De Sa, and C. Ré, "TART: A plug-and-play Transformer module for task-agnostic reasoning," Jun. 2023, arXiv:2306.07536 [cs]. [Online]. Available: <http://arxiv.org/abs/2306.07536>
- [23] A. Bandura, "Social cognitive theory of mass communication," *Media Psychology*, vol. 3, pp. 265–299, 2001.
- [24] B. Fogg, "A behavior model for persuasive design," in *Proceedings of the 4th International Conference on Persuasive Technology*, ser. Persuasive '09. New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 1–7. [Online]. Available: <https://doi.org/10.1145/1541948.1541999>
- [25] J. Fredricks, *Eight Myths of Student Disengagement: Creating Classrooms of Deep Learning*. Thousand Oaks, California: Corwin Press, 2014. [Online]. Available: <https://sk.sagepub.com/books/eight-myths-of-student-disengagement>
- [26] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling Instruction-Finetuned Language Models," Dec. 2022, arXiv:2210.11416 [cs]. [Online]. Available: <http://arxiv.org/abs/2210.11416>
- [27] 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]. [Online]. Available: <http://arxiv.org/abs/2109.01652>
- [28] X. Liu, "Methods for handling missing data," in *Methods and Applications of Longitudinal Data Analysis*, X. Liu, Ed. Academic Press, 2016, ch. 14, pp. 441–473.
- [29] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert systems with applications*, vol. 73, pp. 220–239, 2017.
- [30] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, 1 2022.
- [31] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.
- [32] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023.
- [33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RcQY7>