Improving Transfer Learning for Early Forecasting of Academic Performance by Contextualizing Language Models

Ahatsham Hayat[1], **Bilal Khan**[2], **Mohammad Rashedul Hasan**[1]
University of Nebraska-Lincoln[1], Lehigh University[2]
aahatsham2@huskers.unl.edu, bik221@lehigh.edu, hasan@unl.edu

Abstract

This paper presents a cutting-edge method that harnesses contextualized language models (LMs) to significantly enhance the prediction of early academic performance in STEM fields. Our approach uniquely tackles the challenge of transfer learning with limited-domain data. Specifically, we overcome this challenge by contextualizing students' cognitive trajectory data through the integration of both distal background factors (comprising academic information, demographic details, and socioeconomic indicators) and proximal non-cognitive factors (such as emotional engagement). By tapping into the rich prior knowledge encoded within pre-trained LMs, we effectively reframe academic performance forecasting as a task ideally suited for natural language processing.

Our research rigorously examines three key aspects: the impact of data contextualization on prediction improvement, the effectiveness of our approach compared to traditional numeric-based models, and the influence of LM capacity on prediction accuracy. The results underscore the significant advantages of utilizing larger LMs with contextualized inputs, representing a notable advancement in the precision of early performance forecasts. These findings emphasize the importance of employing contextualized LMs to enhance artificial intelligence-driven educational support systems and overcome data scarcity challenges.

1 Introduction

Modern artificial intelligence (AI) methods, such as deep learning (DL), have increasingly been deployed as cost-effective solutions to develop early-warning systems across various sectors, including health (Adler et al., 2022; Mamun et al., 2022; Zhao et al., 2019; Horwitz et al., 2022; Liu et al., 2023a; Collins et al., 2023; Xu et al., 2023; Adler et al., 2020) and education (Wang et al., 2016, 2014; Li et al., 2020; Xu and Ouyang, 2022). These systems leverage forecasting-based interventions to preemptively address potential issues, from medical conditions to academic performance. In the educational domain, specifically, AI-based interventions utilize cognitive data, like students' course-related assessment scores, to predict and improve academic outcomes (Greenstein et al., 2021; Arnold and Pistilli, 2012; Liu et al., 2023b). The efficacy of these interventions hinges on the precision of early forecasts—predicting course performance as early as possible (Hasan and Aly, 2019; Hasan and Khan, 2023). However, this poses a significant challenge when training data is scarce, leading to suboptimal model performance. Transfer learning could offer a solution, yet the approach is hampered by the lack of relevant pre-trained models or sufficiently large, domain-specific datasets for pre-training (Tsiakmaki et al., 2020).

In this paper, we address the challenges associated with limited training data by introducing a novel transfer learning methodology specifically tailored for domain-specific data within STEM (Science, Technology, Engineering, and Mathematics) education contexts. We propose leveraging Transformer-based (Vaswani et al., 2017) pre-trained language models (LMs) for early prediction of academic performance in undergraduate STEM courses. Our method exploits the extensive knowledge base (Raffel et al., 2020; Roberts et al., 2020) and reasoning capabilities (Chowdhery et al., 2022; Wei et al., 2023; Bhatia et al., 2023) of LMs, transforming end-of-the-semester performance forecasting into a natural language text generation task.

To enhance knowledge transfer using limited domain data, we **contextualize** students' cognitive data by integrating both distal background factors and proximal non-cognitive factors. This multi-dimensional approach encompasses demographic, socioeconomic, and academic background factors, as well as non-cognitive features like emotional engagement, to enrich the predictive model. By

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 137–148

transforming the ordinal (numeric or real-valued) features of our data into natural language text sequences, we tailor pre-trained LMs to our specific task. Additionally, we augment these sequences to increase the dataset size, thereby improving predictive accuracy through a more balanced representation of various performance outcomes.

Contextualizing Academic Trajectories. Our approach integrates students' background and engagement data to provide a comprehensive view of their academic journey. Based on Social Cognitive Career Theory (Bandura, 2001), we hypothesize that a student's course performance correlates with their background, suggesting that LMs can learn individualized academic patterns. Furthermore, longitudinal non-cognitive data, reflecting aspects like motivation and engagement, are posited to have a strong correlation with students' academic trajectories, potentially enhancing the LMs' predictive accuracy (Fogg, 2009; Fredricks, 2014).

Our contextualization process divides into four categories:

- **Demographic Contextualization**: Includes inherent personal and social identity factors, such as race and gender. These are critical for understanding the diverse identities students bring to their educational experiences and how these aspects influence their academic outcomes in the course.
- **Socioeconomic Contextualization**: Encompasses factors related to the economic status and background of the student's family, like parent's total yearly income. This contextualization helps to understand the resources and socio-economic pressures that might influence a student's academic performance and opportunities.
- **Academic Contextualization**: Pertains to the specifics of a student's educational path, including their class standing year (freshman, sophomore, junior, senior) and their chosen major. This type of contextualization is vital for understanding how students' educational choices and progression affect performance.
- **Emotional Engagement Contextualization**: Centers on students' emotional and perceptual dimensions of academic engagement. Specifically, it aims to explore how students' anticipations of academic outcomes (expected grades) and their satisfaction with their academic performance influence their engagement, motivation, and overall educational journey.

Using the contextualized academic trajectory data, we address the following research questions.

- **[RQ1]**: How does contextualization of academic trajectory data impact the efficacy of transfer learning from pre-trained LMs in early academic performance forecasting?
- **[RQ2]**: How does a natural language text generation approach compare with numeric feature-based models in early performance forecasting?
- **[RQ3]**: What impact does the capacity of pre-trained LMs (i.e., the number of parameters) have on forecasting accuracy?

Our primary contributions are threefold.

Innovative Methodology: We introduce a novel methodology that employs natural language text generation for the early forecasting of academic performance, showcasing a unique blend of linguistic and educational insights.

Contextualization as a Catalyst for Transfer Learning: We demonstrate that contextualizing academic trajectory data significantly enhances the transfer learning process from pre-trained LMs. By embedding both cognitive and non-cognitive features within a rich contextual narrative, our approach unlocks the vast potential of LMs to understand and predict academic outcomes with remarkable accuracy.

Exploitation of Pre-trained LM Knowledge: Our research underscores the pivotal role of leveraging the inherent, comprehensive knowledge encapsulated within LMs. Through our method, we illustrate how the nuanced understanding and versatility of LMs can be effectively harnessed for the domain-specific task of predicting student performance, thus marking a significant advancement in the field of educational AI.

The remainder of the paper is organized as follows: Section 2 outlines our methodology, encompassing a description of the dataset and its collection. In Section 3, we present the experiments and provide a detailed analysis of the results, followed by our conclusions and suggestions for future work in Section 4. Finally, Section 5 offers a discussion of pertinent literature.

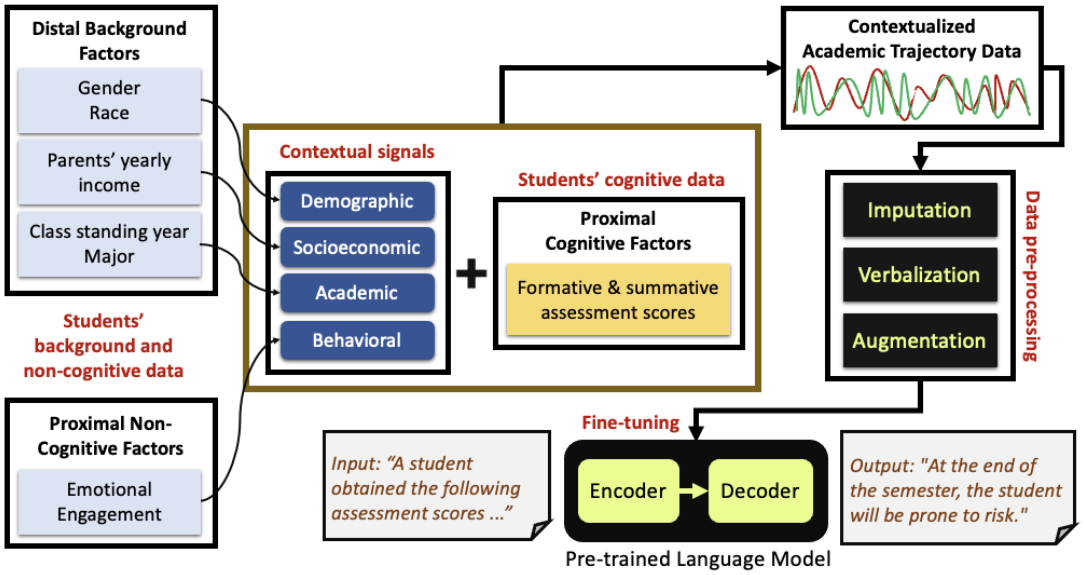


Figure 1: An overview of the approach for enhancing transfer learning from pre-trained language models for early academic performance forecasting.

2 Method

To harness the nuanced understanding pre-trained LMs offer regarding students' academic experiences, we assembled a **detailed longitudinal dataset** that examines the interplay among various factors, including background, cognitive, and non-cognitive elements in student learning. Figure 1 illustrates the LM-based transfer learning framework, featuring the contextualization of proximal cognitive data followed by the preprocessing of the contextualized academic trajectory. Data contextualization involves integrating distal background and proximal non-cognitive factors with cognitive trajectory data. Below, we outline the process of compiling a language dataset, encompassing data collection and pre-processing methods, and conclude with a formal description of transfer learning through fine-tuning of LMs.

2.1 Data Collection

Our dataset comprises information obtained from 48 first-year college students enrolled in an introductory programming course at a public university in the United States, following approval from the University's Institutional Review Board. The dataset encompasses three key dimensions of the students' academic journeys.

Background Data (5-dimensional): At the outset of the semester, critical 5-dimensional background data was collected through a Qualtrics-based multiple-choice web survey. This numeric dataset includes students' academic details (such as class standing year and major), demographic information (including gender and race), and a socioeconomic indicator (family yearly income).

Non-Cognitive Data (2-dimensional): This dimension includes longitudinal measures of students' emotional engagement throughout the semester, comprising 2-dimensional data reflecting students' anticipated end-of-semester performance and their current performance satisfaction, both in numeric format.

The data is collected via a **privacy-preserving smartphone application**, designed to prompt contextually relevant, study-specific multiple-choice questions daily. This ensures that participants' anonymized responses are securely compiled on cloud servers for subsequent analysis. Each participant is assigned a unique randomly generated ID upon enrollment, with no personally identifiable information collected via the app. All data collected is tagged solely by the participant's random ID, with no linkage maintained between the ID and participant identity. Geolocation and Bluetooth sensors are utilized in the app to ascertain instantaneous context for question triggers, although sensor data is not persistently stored. By transparently informing students about the privacy-preservation mechanisms, we mitigate potential psychological and academic incentives for artificial performance or dishonest responses during experience sampling.

Furthermore, this privacy-preserving mechanism serves to mitigate potential biases in the data collection process. By anonymizing participants' responses and ensuring that no personally identifiable information is collected, we minimize the risk of participants feeling pressured to provide socially desirable responses. This approach promotes more authentic and unbiased data collection, contributing to the reliability and validity of our findings.

Cognitive Data (21-dimensional): The dataset also includes 21-dimensional numeric cognitive data derived from students' assessment scores (both formative and summative) over the first 8 weeks of the semester. This cognitive data was obtained from the course's learning management system, Canvas, providing insights into students' academic performance, engagement, and progress within the course curriculum.

2.2 Data Contextualization

We enriched students' cognitive trajectory data—comprising their course-related formative and summative scores—by incorporating four contextual dimensions: demographic (gender and race), academic (class standing year and major), socioeconomic (family yearly income), and behavioral (emotional engagement). The dynamic cognitive and non-cognitive data were intertwined to preserve their temporal sequence, while the static background data was added at the end of the trajectory.

2.3 Data Pre-processing

The contextualized numeric trajectory data underwent preprocessing to adapt it for LM use, which included handling missing values in the non-cognitive data, verbalization of the data, and data augmentation for enhanced model training.

Data Imputation. The proximal non-cognitive data exhibited missing values, resulting from participants either skipping questions or temporarily uninstalling the app. We encountered two distinct patterns of missing data: complete absence of responses for an entire day and partial absences within a day. To address days with entirely missing data, we employed the Last Observation Carried Forward (LOCF) imputation method (Liu, 2016). This method involves carrying forward the last observed value for each participant to replace missing values at subsequent time points. While LOCF is a commonly used approach due to its simplicity, it assumes that the missing data points would have followed a similar trajectory as the last observed value. In situations where no prior data were available, the Next Observation Carried Backward (NOCB) approach was employed (Jahangiri et al., 2023), using data from a subsequent day that contained relevant responses. The challenge of partially missing data, particularly for follow-up questions, necessitated a more nuanced approach. When the preceding day's trigger question response did not match, directly applying LOCF for the follow-up question was deemed unreliable (Lachin, 2016). Instead, we filled these gaps with responses from days where the trigger question responses aligned. If no matching previous day could be identified, a future day with corresponding answers was utilized.

Data Verbalization. To transform the numeric dataset into natural language, we designed a template for verbalizing both the input (X) and output (Y) data sequences (refer to the Appendix for details). Input sequences were prefaced with contextual messages, such as "A student obtained the following assessment scores in an introductory programming course ..." for cognitive data, and "Some background information about the student: ..." for distal data. Chronological order was emphasized by prefacing data with the week number, e.g., "*In week [WEEK_NUMBER]*". The output sequences, categorized into four performance groups (at-risk, prone-to-risk, average, outstanding), contextualized the final letter grade in a natural language expression, e.g., "*At the end of the semester, the student will be at risk.*" . This verbalization process yielded three datasets based on 8-week, 4-week, and 2-week long input sequences.

Data Augmentation. Given the initial dataset's unbalanced distribution across performance categories (24 instances of outstanding, 12 average, 6 prone-to-risk, and 6 at-risk), we employed a two-fold approach for data augmentation. Firstly, we utilized oversampling techniques (Haixiang et al., 2017; Hernandez et al., 2013) to duplicate instances from minority classes, thus balancing the dataset. Secondly, we incorporated synonym replacement methods (Li et al., 2022), which involved substituting words with their synonyms to introduce token variations. This comprehensive approach aimed to not only address class imbalance but also enrich the dataset with diverse token variations.

As a result of our data augmentation strategy, the augmented dataset showcased a more equitable dis-

tribution among performance categories, totaling 144 samples, comprising 48 instances of outstanding, 36 average, 30 prone-to-risk, and 30 at-risk.

These methodologies provide a robust foundation for applying transfer learning to LMs, facilitating a deep understanding of students' academic performance through a multi-dimensional data lens.

2.4 Fine-tuning LMs

Each sequence in X and Y contains standard lexical literals used in English (e.g., words and phrases), which is used to fine-tune a pre-trained encoder-decoder LM. The encoder $f_E(.)$ maps the input sequence $(x_1, x_2, ..., x_l)$ to an intermediate latent embedding sequence $(z_1, z_2, ..., z_l)$.

$$z = f_E(x_1, x_2, ..., x_l; \theta_E) \quad (1)$$

where θ_E are the weights of the encoder.

The decoder $f_D(.)$ takes the latent embeddings $(z_1, z_2, ..., z_l)$ to generate an output sequence $(\hat{y}_1, \hat{y}_2, ..., \hat{y}_m)$ in an auto-regressive fashion, i.e., at each step the decoder $f_D(.)$ uses previously generated symbols $\hat{y}_{<m}$ as additional input for generating the next token $\hat{y}_m$.The probability of generating the m-th token $\hat{y}_m$ is given by

$$p(\hat{y}_m|\hat{y}_{<m}; z_1, z_2, ..., z_l) = softmax(f_D(\hat{y}_{<m}; z_1, z_2, ..., z_l; \theta_D)) \quad (2)$$

where θ_D are the weights of the decoder. For fine-tuning the encoder-decoder LM, the multi-class cross-entropy loss function is used. The number of classes in the loss function is set by the total number of tokens in the vocabulary. For a batch size B, the loss function is:

$$\mathcal{L} = -\sum_{b=1}^{B}\sum_{m=1}^{M} y_m^b log\hat{y}_m^b \quad (3)$$

3 Experiments

To thoroughly investigate the research questions outlined in Section 1, we performed a series of experiments focusing on the learning capabilities of LMs. These experiments involved fine-tuning pre-trained LMs across multi-dimensional language datasets spanning 8 weeks, 4 weeks, and 2 weeks. This selection of timeframes facilitated an in-depth examination of LM adaptability over various periods. The effectiveness of the adapted LMs was assessed through their ability to identify performance types based on matching keywords in the predicted output sequences. Moreover, we explored the impact of LM size—small, medium, and large—on their performance.

Experimental Setup. For the encoder-decoder LM, we used pre-trained FLAN-T5 (Chung et al., 2022), which is a variant of the T5 model (Raffel et al., 2020). The FLAN-T5 model is instruction fine-tuned, making it suitable for our purposes. We employed FLAN-T5 with three different capacities, determined by the number of parameters: FLAN-T5-Small (80M), FLAN-T5-Base (250M), and FLAN-T5-Large (770M). These LMs have a context window limited to 512 tokens. As baseline comparisons, we utilized four models that work with only numeric features: three neural networks (NNs) and one non-NN machine learning model. The neural networks include a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), a Convolutional Neural Network (CNN) with a one-dimensional (1D) convolutional kernel (Kim, 2014), and a Transformer network (Vaswani et al., 2017). The non-NN machine learning model employed was a Support Vector Machine (SVM) with a linear kernel (Boser et al., 1992), which demonstrated superior performance over the Gaussian Radial Basis Function kernel.

The baseline models were trained using 3 variably-length numeric datasets containing only the cognitive features. Exploring baseline models with all three feature types is planned as future work. To ensure compatibility with the LM-based experiments, the numeric datasets were created from the augmented verbalized datasets by decoding the cognitive feature part of text sequences into numeric values.

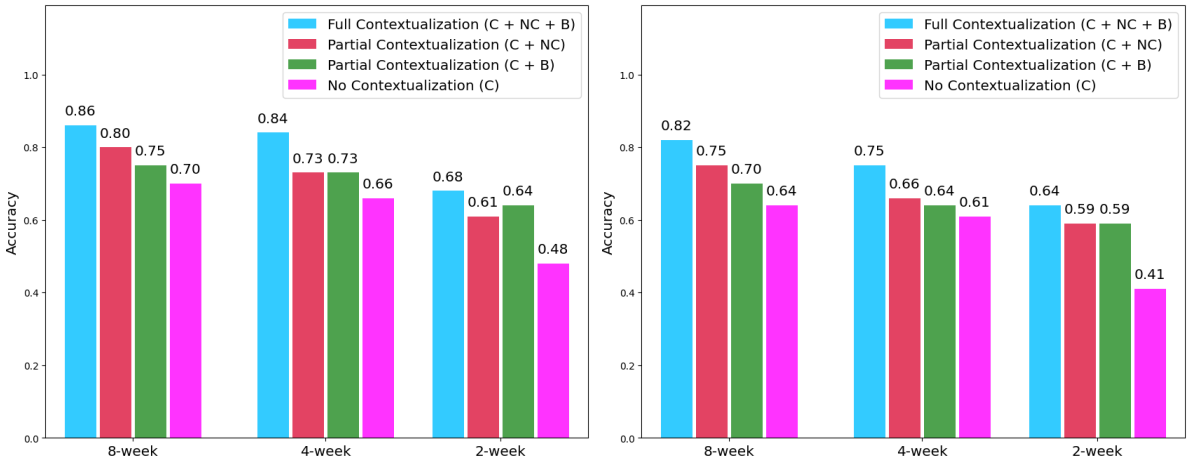
We used the same test sets to evaluate both model types, employing the following metrics: accuracy, precision, recall, and F1 score. A detailed description of the experimental setup is provided in the Appendix.

3.1 Results

[RQ1]: *How does contextualization of academic trajectory data impact the efficacy of transfer learning from pre-trained LMs in early academic performance forecasting?* The core objective of this study is to evaluate how the contextualization of academic trajectory data influences the forecasting effectiveness of pre-trained LMs. To this end, we fine-tuned LMs of varying sizes with aca-

Table 1: Evaluation of the large LM (FLAN-T5-Large) fine-tuned with four combinations of the 3 feature types using the 8-week, 4-week, and 2-week datasets. The best results are in **bold**.
Legends: C=Cognitive, NC=Non-Cognitive, B=Background, AR=At-Risk, PR=Prone-To-Risk, AV=Average, OU=Outstanding, P=Precision, R=Recall, F1=F1 Score, A=Accuracy

Features	Class	8-week				4-week				2-week			
		P	R	F1	A	P	R	F1	A	P	R	F1	A
Full Contextualization (C + NC + B)	AR	0.78	1.00	0.88	**0.89**	1.00	1.00	1.00	**0.84**	0.64	1.00	0.78	**0.77**
	PR	0.89	0.80	0.84		0.89	0.80	0.84		1.00	0.50	0.67	
	AV	0.92	1.00	0.96		0.71	0.91	0.80		0.73	1.00	0.85	
	OU	0.93	0.81	0.87		0.86	0.75	0.80		0.85	0.69	0.76	
Partial Contextualization (C + NC)	AR	0.70	1.00	0.82	0.82	0.70	1.00	0.82	0.77	0.62	0.71	0.67	0.68
	PR	1.00	0.60	0.75		0.86	0.60	0.71		0.71	0.50	0.59	
	AV	0.73	1.00	0.85		0.69	1.00	0.81		0.62	0.91	0.74	
	OU	0.92	0.75	0.83		0.91	0.62	0.74		0.77	0.62	0.69	
Partial Contextualization (C + B)	AR	0.78	1.00	0.88	0.77	0.88	1.00	0.93	0.77	0.60	0.86	0.71	0.64
	PR	0.89	0.80	0.84		0.71	1.00	0.83		0.71	0.50	0.59	
	AV	0.67	0.73	0.70		0.69	0.82	0.75		0.70	0.64	0.67	
	OU	0.79	0.69	0.73		0.89	0.50	0.64		0.59	0.62	0.61	
No Contextualization (C)	AR	0.60	0.86	0.71	0.73	0.62	0.71	0.67	0.70	0.36	0.57	0.44	0.52
	PR	0.86	0.60	0.71		0.67	0.60	0.63		0.88	0.70	0.78	
	AV	0.60	0.82	0.69		0.67	0.91	0.77		0.54	0.64	0.58	
	OU	0.92	0.69	0.79		0.83	0.62	0.71		0.42	0.31	0.36	

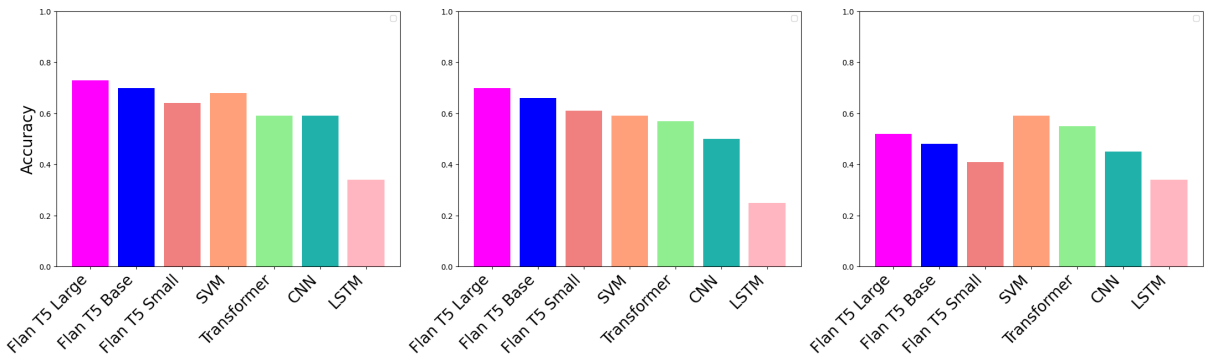


(a) FLAN-T5 Base (b) FLAN-T5 Small

Figure 2: Impact of contextualization on the FLAN-T5 Base and Small models.

demic trajectory data enriched with three types of features: cognitive (C), non-cognitive (NC), and background (B). This investigation includes comparing the performance impact between fully contextualized LMs (utilizing all three feature types) and partially-contextualized or non-contextualized LMs. For partial contextualization, we explored combinations of C+NC and C+B features, whereas, in the non-contextualization scenario, only cognitive (C) features were employed for model fine-tuning.

According to the performance metrics provided in Table 1 for the best-performing large LM, FLAN-T5-Large, it is evident that models utilizing a contextualization approach, whether fully or partially, significantly outperform those without any contextualization. Specifically, the **fully contextualized LMs demonstrate superior forecasting abilities**. For instance, such a model can predict student performance with an accuracy of 77% by the end of the 2nd week of the semester. This early prediction capability is vital for implementing effective early



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

Figure 3: Comparison with baseline models on cognitive features.

Table 2: Evaluation of the three baseline models trained with cognitive features using the 8-week, 4-week, and 2-week datasets. The best results are in **bold**.
Legends: AR=At-Risk, PR=Prone-To-Risk, AV=Average, OU=Outstanding, P=Precision, R=Recall, F1=F1 Score, A=Accuracy

Model	Class	8-week				4-week				2-week			
		P	R	F1	A	P	R	F1	A	P	R	F1	A
CNN	AR	0.50	0.86	0.63	0.59	0.44	0.57	0.50	0.50	0.45	0.71	0.56	0.45
	PR	0.83	0.50	0.62		1.00	0.30	0.46		0.44	0.70	0.54	
	AV	1.00	0.09	0.17		0.33	0.55	0.43		0.22	0.18	0.20	
	OU	0.56	0.88	0.68		0.37	0.56	0.58		0.75	0.38	0.50	
LSTM	AR	1.00	0.14	0.25	0.34	0.00	0.00	0.00	0.25	0.15	0.29	0.20	0.34
	PR	0.27	0.40	0.32		0.00	0.00	0.00		0.00	0.00	0.00	
	AV	0.33	0.27	0.30		0.26	0.73	0.38		0.00	0.00	0.00	
	OU	0.37	0.44	0.40		0.33	0.19	0.24		0.42	0.81	0.55	
Transformer	AR	0.78	1.00	0.88	0.59	0.54	1.00	0.70	0.57	0.56	0.71	0.63	0.55
	PR	0.57	0.40	0.47		1.00	0.60	0.75		0.80	0.60	0.71	
	AV	0.41	0.64	0.50		0.40	0.18	0.25		0.00	0.00	0.00	
	OU	0.73	0.50	0.59		0.50	0.62	0.56		0.46	0.81	0.59	
SVM	AR	1.00	0.71	0.83	**0.68**	1.00	0.86	0.92	**0.59**	0.54	0.78	0.64	**0.59**
	PR	0.88	0.78	0.82		1.00	0.33	0.50		1.00	0.20	0.33	
	AV	0.41	0.88	0.56		0.38	0.38	0.38		0.67	0.50	0.57	
	OU	0.67	0.46	0.55		0.38	0.62	0.47		0.57	0.76	0.65	

intervention strategies.

Moreover, identifying students at risk (AR) or prone to risk (PR) early is crucial for timely support. The 2-week model, when fully contextualized, exhibits a remarkable recall rate of 100% for the AR group. As more data becomes available, the 4-week model maintains this 100% recall for the AR group and also achieves an 80% recall for the PR group, both of which are essential for early intervention efficacy. Expanding the data window to 8 weeks further enhances the model's accuracy to 89%, underlining the benefits of full contextualization in improving early detection and intervention outcomes.

Partial Contextualization was explored in two variations: one combining cognitive and non-cognitive features (C + NC) and the other cognitive and background features (C + B). The C + NC configuration demonstrated moderate success, with overall accuracy ranging from 68% to 82%, indicating a somewhat effective use of student information minus the background context. In contrast, the C + B setup, omitting non-cognitive traits, showed a slight decrease in performance, particularly for the 2-week predictions, where accuracy dropped to 64%. These outcomes highlight the nuanced contribution of non-cognitive factors in short-term risk assessment.

No Contextualization (C alone) presented the **most significant drop in performance**, with ac-

curacy falling to 52% for the 2-week predictions. This stark decrease underscores the critical role of contextualization in enhancing the predictive power of the model.

In addressing RQ1, the evaluation of the FLAN T5 Base model also underscores the importance of academic trajectory data contextualization (see Figure 2(a)). When fine-tuned with a comprehensive set of features (C + NC + B), it demonstrates a clear advantage, achieving accuracies of 86%, 84%, and 68% across 8-week, 4-week, and 2-week forecasts, respectively. This trend highlights the efficacy of full contextualization in enhancing model performance, despite a slight performance dip compared to the larger model variant, affirming the significance of a rich feature set for improved predictive accuracy.

The investigation with the FLAN T5 Small model further supports the value of contextualization (see Figure 2(b)), achieving peak accuracies of 82%, 75%, and 64% across the same timeframes with full feature integration. Despite facing challenges in short-term risk prediction, the Small model's performance emphasizes the critical role of a comprehensive feature blend in maintaining predictive accuracy, even with constrained computational resources. These findings collectively validate that full contextualization substantially benefits the forecasting capabilities of pre-trained LMs across different model sizes.

[RQ2]: ***How does natural language text generation compare to numeric feature-based models in forecasting early academic performance, using only cognitive features?*** Our analysis contrasts the efficacy of three varying-capacity LMs against four numeric feature-based baseline models, focusing solely on the cognitive features of our dataset. As illustrated in Figure 3 for datasets spanning 8-week, 4-week, and 2-week intervals, the results demonstrate distinct performance dynamics. In the 4-week and 8-week forecasts, LMs consistently outperform the numeric baseline models. Yet, in the initial 2-week forecast, numeric models, specifically the SVM and Transformer, with accuracies of 59% and 55% respectively, outdo the large LM, which records a 52% accuracy. Remarkably, the SVM's performance plateaus at 59% accuracy for the 4-week datasets, in contrast to the large LM, which notably enhances its accuracy to over 70% consistently across the 4-week duration. Detailed comparisons of baseline model performances are provided in Table 2.

[RQ3]: ***What impact does the capacity of pre-trained LMs (i.e., the number of parameters) have on forecasting accuracy?*** Analyzing the test accuracies among the three differently sized LMs (refer to Table 1, Figures 2 and 3) reveals a clear trend: larger models demonstrate enhanced forecasting capabilities. Notably, even after implementing full contextualization, the recall for the at-risk group in the smaller and medium-sized models stands at 86%, while the large model achieves a recall of 100%. This pattern strongly indicates that **achieving optimal early forecasting through the contextualization of LMs is more effective with the deployment of large language models (LLMs)**.

4 Conclusion

In this paper, we ventured into the realm of leveraging modern AI, particularly deep learning and transfer learning methodologies, to tackle the critical challenge of early performance forecasting in the educational sector. Our investigation centered on the innovative use of Transformer-based pre-trained LMs for predicting undergraduate STEM course outcomes, marking a significant departure from traditional numeric feature-based models. By integrating a novel transfer learning approach tailored for small-domain data within STEM education, we aimed to overcome the limitations posed by sparse training datasets, a common hurdle in the educational domain.

Our methodology hinged on the contextualization of academic trajectory data, incorporating a rich tapestry of both cognitive and non-cognitive factors. Through this multi-dimensional approach, we enhanced the LMs' capacity to understand and predict academic performance, achieving a notable improvement in forecasting accuracy. Specifically, we demonstrated that:

- Contextualizing academic trajectory data significantly enhances the transfer learning process from pre-trained LMs, as evidenced by our responses to [RQ1].
- Compared to numeric feature-based models, our natural language text generation approach shows superior performance in early academic forecasting, addressing [RQ2].
- The capacity of pre-trained LMs, in terms of their number of parameters, plays a crucial

role in forecasting accuracy, with larger models outperforming their smaller counterparts, as explored in [RQ3].

These insights underscore the transformative potential of AI-driven tools in proactively identifying and supporting students at risk, thereby enhancing educational outcomes. By leveraging the vast knowledge encapsulated within LMs and enriching it with detailed contextual data across demographic, socioeconomic, academic, and emotional engagement dimensions, we not only tailored the pre-trained LMs to our specific task but also enriched the predictive model with a comprehensive understanding of students' academic journeys.

Looking ahead, our work opens the door to future research in several key areas. Integrating more detailed contextual signals such as real-time academic engagement and behavioral data could enhance LM predictive accuracy, leveraging advances in natural language processing and sentiment analysis to understand students' emotional and cognitive states better. Expanding our approach to a wider range of educational contexts and disciplines would help validate its scalability and adaptability. Additionally, exploring continual learning techniques for LMs might illuminate how to improve forecasting systems' accuracy and reliability over time without extensive retraining. Addressing the ethical and privacy concerns inherent in using detailed student data is also crucial, necessitating robust data governance and ethical AI frameworks to protect students' rights and ensure equitable benefits.

5 Related Work

In advancing educational forecasting, we introduce a distinct approach by applying transfer learning from pre-trained LMs to contextualized time-series data of academic trajectories. This dataset uniquely incorporates both cognitive and non-cognitive features, enriching the forecasting model with a detailed temporal perspective.

Research in time-series forecasting with pre-trained LMs splits into two main streams: data-centric and model-centric approaches (Sun et al., 2023). **Data-centric** methods focus on transforming time-series data into formats amenable to LMs, employing innovative embedding techniques to match time-series data with the textual embedding space of LMs. These techniques range from embedding alignment and augmentation (Sun et al., 2023) to two-stage fine-tuning (Chang et al., 2023) and zero-shot preprocessing for numerical data (Gruver et al., 2023). **Model-centric** strategies, on the other hand, adapt pre-trained LMs specifically for time-series forecasting. This involves fine-tuning certain LM components while introducing time series-specific modifications such as decomposition and soft prompts (Cao et al., 2023), aiming to formulate forecasting as a question-answering task (Xue and Salim, 2023), and prompt-tuning with few-shot learning (prompt engineering) (Liu et al., 2023c).

Our work diverges by leveraging a model-centric approach tailored to the contextual data of academic paths, utilizing discrete prompts. This novel strategy emphasizes the importance of transfer learning from pre-trained LMs to enrich forecasting with a deep, context-aware analysis, setting our research apart in the field of educational forecasting.

6 Limitations

Our study has made important progress in showing how contextualized language models (LMs) can predict early academic performance. Yet, we must acknowledge some limitations that define our research's scope and point towards future research directions.

Data Scope and Diversity: The primary focus of our research on undergraduate STEM courses may circumscribe the applicability of our findings across different academic disciplines and educational levels. The distinct cognitive and engagement challenges inherent to non-STEM subjects underscore the need for subsequent studies aimed at adapting and validating our methodology in a wider educational context.

Model Size and Computational Resources: The deployment of LMs brings to the fore the exigencies of computational resources. The high computational overhead required for the training and operational deployment of these models might preclude their adoption in institutions with limited technological infrastructure, potentially curtailing the broad-scale application of our approach in varied educational settings.

Ethical and Privacy Concerns: Leveraging detailed personal and contextual data of students necessitates a careful navigation of ethical and privacy considerations. While our study has endeavored to

adhere to these imperatives scrupulously, the expansive use of similar methodologies demands a rigorous commitment to data protection standards and ethical practices to mitigate the risk of infringing upon student privacy.

Temporal Dynamics: Our forecasting approach captures a static slice of contextual data, possibly overlooking the dynamic nature of student engagement and performance, which are subject to change over the academic term. The challenge of incorporating continuous data updates into LMs without necessitating extensive retraining poses a significant question for future research.

Interpretability and Explainability: The opaque nature of LMs, as with many deep learning models, presents a barrier to interpretability and explainability. To engender trust among educational practitioners and stakeholders, it is imperative to develop methodologies that elucidate the rationales behind model predictions in a comprehensible manner.

Bias and Fairness: The risk of propagating biases through pre-trained LMs, a reflection of their training datasets, is a critical concern. These biases have the potential to skew forecasting accuracy and fairness, impacting various student demographics disparately. Vigilance to prevent the reinforcement of existing educational disparities is essential.

Computational Limitations: Our investigation's scope was notably constrained by the limited memory capacity of available GPUs. This limitation thwarted our ability to fully leverage the spectrum of distal and proximal non-cognitive features, employ rich and expressive instructional prompts, and utilize LMs with $\geq$ 1 billion parameters. Overcoming these computational hurdles is crucial for unlocking the full potential of LLMs in educational forecasting.

These limitations underscore the imperative for continued research to surmount these hurdles. Future endeavors should focus on broadening the inclusivity, ethical integrity, and scalability of AI-driven educational interventions, ensuring they serve as equitable and effective support mechanisms across the diverse landscape of learning environments.

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

Daniel A Adler, Dror Ben-Zeev, Vincent W-S Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. 2020. Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks. *JMIR mHealth and uHealth*, 8(8):e19962.

Daniel A. Adler, Fei Wang, David C. Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE*, 17(4):e0266516. Publisher: Public Library of Science.

Kimberly E. Arnold and Matthew D. Pistilli. 2012. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270.

A. Bandura. 2001. Social cognitive theory of mass communication. *Media Psychology*, 3:265–299.

Kush Bhatia, Avanika Narayan, Christopher De Sa, and Christopher Ré. 2023. TART: A plug-and-play Transformer module for task-agnostic reasoning. ArXiv:2306.07536 [cs].

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, pages 144–152, Pittsburgh, PA, USA. ACM Press.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting.

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob

Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. ArXiv:2210.11416 [cs].

Amanda C. Collins, Damien Lekkas, Matthew David Nemesure, Tess Z. Griffin, George Price, Arvind Pillai, Subigya Nepal, Michael V. Heinz, Andrew T. Campbell, and Nicholas C. Jacobson. 2023. Semantic signals in self-reference: The detection and prediction of depressive symptoms from the daily diary entries of a sample with major depressive disorder.

Nello Cristianini and John Shawe-Taylor. 1999. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*. Cambridge University Press, USA.

BJ Fogg. 2009. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Jennifer Fredricks. 2014. *Eight Myths of Student Disengagement: Creating Classrooms of Deep Learning*. Corwin Press, Thousand Oaks, California.

Nathan Greenstein, Grant Crider-Phillips, Claire Matese, and Sung-Woo Cho. 2021. Predicting Student Outcomes to Drive Proactive Support: An Exploration of Machine Learning to Advance Student Equity & Success. Technical report, University of Oregon.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239.

Mohammad Hasan and Bilal Khan. 2023. A Trajectory-Clustering Framework for Assessing AI-Based Adaptive Interventions in Undergraduate STEM Learning American Society for Engineering Education.

Mohammad Rashedul Hasan and Mohamed Aly. 2019. 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*. Event-place: Las Vegas, Nevada.

Julio Noe Hernandez, Jesús Ariel Carrasco-Ochoa, and José Francisco Martínez Trinidad. 2013. An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I*, volume 8258 of *Lecture Notes in Computer Science*, pages 262–269. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Adam G. Horwitz, Shane D. Kentopp, Jennifer Cleary, Katherine Ross, Zhenke Wu, Srijan Sen, and Ewa K. Czyz. 2022. Using machine learning with intensive longitudinal data to predict depression and suicidal ideation among medical interns over time. *Psychological Medicine*, pages 1–8.

M. Jahangiri, A. Kazemnejad, K. S. Goldfeld, M. S. Daneshpour, S. Mostafaei, D. Khalili, M. R. Moghadas, and M. Akbarzadeh. 2023. A wide range of missing imputation approaches in longitudinal data: a simulation study and real data analysis. *BMC Med Res Methodol*, 23(1):161.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

John M. Lachin. 2016. Fallacies of last observation carried forward analyses. *Clinical trials*, 13(2):161–168.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Xiang Li, Xinning Zhu, Xiaoying Zhu, Yang Ji, and Xiaosheng Tang. 2020. Student Academic Performance Prediction Using Deep Multi-source Behavior Sequential Network. In *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 567–579, Cham. Springer International Publishing.

Haihong Liu, Xiaolei Zhang, Haining Liu, and Sheau Tsuey Chong. 2023a. Using Machine Learning to Predict Cognitive Impairment Among Middle-Aged and Older Chinese: A Longitudinal Study. *International Journal of Public Health*, 68:1605322.

Lydia T. Liu, Serena Wang, Tolani Britton, and Rediet Abebe. 2023b. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, 120(9):e2204781120. Publisher: Proceedings of the National Academy of Sciences.

Xian Liu. 2016. Methods for handling missing data. In Xian Liu, editor, *Methods and Applications of Longitudinal Data Analysis*, chapter 14, pages 441–473. Academic Press.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023c. Large language models are few-shot health learners.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Abdullah Mamun, Krista S. Leonard, Matthew P. Buman, and Hassan Ghasemzadeh. 2022. Multimodal Time-Series Activity Forecasting for Adaptive Lifestyle Intervention Design. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. ISSN: 2376-8894.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. 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)*, pages 5418–5426, Online. Association for Computational Linguistics.

Chenxi Sun, Yaliang Li, Hongyan Li, and Shenda Hong. 2023. Test: Text prototype aligned embedding to activate llm's ability for time series.

Maria Tsiakmaki, Georgios Kostopoulos, Sotiris Kotsiantis, and Omiros Ragos. 2020. Transfer learning from deep neural networks for predicting student performance. *Applied Sciences*, 10(6).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv:1706.03762 [cs].

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. 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, pages 3–14, New York, NY, USA. Association for Computing Machinery.

Rui Wang, Peilin Hao, Xia Zhou, Andrew T. Campbell, and Gabriella Harari. 2016. SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. *GetMobile: Mobile Computing and Communications*, 19(4):13–17.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Weiqi Xu and Fan Ouyang. 2022. The application of AI technologies in STEM education: a systematic review from 2011 to 2021. *International Journal of STEM Education*, 9(1):59.

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):190:1–190:34.

Hao Xue and Flora D. Salim. 2023. PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting. ArXiv:2210.08964 [cs, math, stat].

Juan Zhao, QiPing Feng, Patrick Wu, Roxana A. Lupu, Russell A. Wilke, Quinn S. Wells, Joshua C. Denny, and Wei-Qi Wei. 2019. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Scientific Reports*, 9:717.