A Three-Tier LLM Framework for Forecasting Student Engagement from Qualitative Longitudinal Data

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

Abstract

Forecasting nuanced shifts in student engagement from longitudinal experiential (LE) data—multi-modal, qualitative trajectories of academic experiences over time—remains challenging due to high dimensionality and missingness. We propose a natural language processing (NLP)-driven framework using large language models (LLMs) to forecast binary engagement levels across four dimensions: Lecture Engagement Disposition, Academic Self-Efficacy, Performance Self-Evaluation, and Academic Identity and Value Perception. Evaluated on 960 trajectories from 96 first-year STEM students, our three-tier approach—LLM-informed imputation to generate textual descriptors for missing-not-at-random (MNAR) patterns, zero-shot feature selection via ensemble voting, and fine-tuned LLMs—processes textual non-cognitive responses. LLMs substantially outperform numeric baselines (e.g., Random Forest, LSTM) by capturing contextual nuances in student responses. Encoder-only LLMs surpass decoder-only variants, highlighting architectural strengths for sparse, qualitative LE data. Our framework advances NLP solutions for modeling student engagement from complex LE data, excelling where traditional methods struggle.

1 Introduction

Transformer-based (Vaswani et al., 2017) large language models (LLMs) have significantly advanced natural language processing (NLP), pushing boundaries in text understanding and generation across diverse applications (Bommasani et al., 2021). Beyond excelling in traditional NLP tasks such as summarization and translation (Zhao et al., 2025), LLMs have demonstrated a remarkable capacity for reasoning over complex, context-rich information, suggesting their potential for analyzing sequential and subjective data (Wei et al., 2022; Touvron et al., 2023). One particularly promising, yet relatively untapped, area for LLM application lies in the analysis of **longitudinal experiential (LE)** data—time-series records capturing individuals' evolving perceptions, emotions, and experiences (Xu et al., 2022).

Within educational contexts, LE data offers a unique and valuable perspective on students' subjective engagement, a well-established predictor of retention and academic achievement (Fredricks, 2014; Sinatra et al., 2015). Despite its richness, the inherent characteristics of LE data, including its qualitative nature, temporal dependencies, and frequent missingness, present substantial computational challenges that often limit the effectiveness of traditional machine learning approaches (Xu et al., 2023). Our research focuses on this underexplored intersection of LLMs and the complexities of LE data analysis in education.

In educational research, LE data systematically gathers real-time, self-reported insights—including emotional responses, shifts in motivation, and the development of self-efficacy—from individuals over time, complementing traditional cognitive assessments (Kolb, 1984; Palmer et al., 2010). Understanding these **non-cognitive (NC)** dimensions can reveal critical engagement patterns predictive of academic outcomes, informing timely interventions (Wang et al., 2014; Li et al., 2020).

Our research is based on a dataset of 28 distinct NC features collected weekly from 96 first-year college STEM (science, technology, engineering, mathematics) students across three semesters (Hayat et al., 2024a,b). These features aimed to capture a comprehensive view of their engagement. However, initial analysis revealed that many of these features suffered from **extreme missingness**, with some having up to 100% unanswered responses. To ensure a more robust analysis, we focused on 10 key qualitative NC features that exhibited a response rate of at least 35%. Our forecasting task specifically targets predicting weekly

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 334–347

binary engagement shifts (positive vs. negative) across four critical dimensions using 4-week historical sequences, where each prediction involves determining whether a student's engagement level in week 5 will exceed their average from weeks 1-4. These 10 features, despite the inherent challenges of qualitative data, form the basis of our investigation into student engagement forecasting.

Initial attempts to forecast engagement by converting the textual responses of these 10 features into numeric values (e.g., via Likert scale encoding) and training traditional machine learning models like Random Forest (Breiman, 2001) and Support Vector Machines (Hearst et al., 1998), as well as time-series models like LSTMs (Hochreiter and Schmidhuber, 1997), yielded poor forecasting performance. Similarly, directly fine-tuning standard decoder-only and encoder-only LLMs on the raw text of these 10 NC features also resulted in suboptimal forecasting accuracy, although showing marginal improvement over the numeric-based models. This suggests that while LLMs possess inherent advantages, directly processing all available qualitative features, even after initial filtering for missingness, can still introduce noise, hindering their ability to effectively discern predictive signals in this specific type of LE data.

The limitations observed with both traditional numeric approaches and direct LLM fine-tuning underscore the need for a more tailored strategy for analyzing this qualitative, time-series LE data with significant missingness. Unlike traditional time-series models (e.g., ARIMA (Box et al., 2015), LSTMs (Hochreiter and Schmidhuber, 1997)), which struggle with non-numeric input and are particularly vulnerable to biases introduced by missing data, LLMs offer the potential to directly process qualitative information.

As highlighted earlier, processing LE data presents a complex array of challenges, including its qualitative nature, temporal dependencies, and significant sparsity due to missing self-reports. These difficulties are further compounded by the prevalence of missing-not-at-random (MNAR) patterns (Rubin, 1976), where the absence of a report is often correlated with the very engagement phenomena we aim to study. This introduces biases that conventional statistical imputation techniques, such as Last Observation Carried Forward (LOCF), are often inadequate to handle effectively (Schafer, 1997).

To address this critical issue of biased missingness and the noise within the qualitative LE feature space, we propose a *three-tier LLM framework* specifically designed for the unique characteristics of this LE data: (1) **LLM-informed imputation**, using LLMs' contextual understanding to generate textual descriptors for missing values, mitigating MNAR bias where traditional methods fall short; (2) **LLM-based zero-shot feature selection**, employing a panel of expert LLMs to infer and select the most relevant subset of our 10 qualitative NC features via majority voting, thereby reducing noise; and (3) **fine-tuned forecasting**, comparing decoder-only and encoder-only LLMs to predict binary engagement levels for *four key dimensions: Lecture Engagement Disposition (LED), Academic Self-Efficacy (ASE), Performance Self-Evaluation (PSE), and Academic Identity and Value Perception (AIVP)*.

Evaluated on 960 overlapping 4-week trajectories (weeks 1-4 predicting week 5) derived from our dataset, our three-tier approach significantly outperforms numeric baselines. Ablation studies further demonstrate the efficacy of each component: (1) zero-shot feature selection yields substantial gains compared to using all 10 NC features, highlighting the noise reduction achieved through expert LLM guidance; and (2) LLM-based feature selection surpasses numeric feature-based models that utilize all available features, directly justifying the need for our LLM-driven feature selection process for this qualitative data. Encoder-only architectures consistently outperform decoder-only variants in this sparse LE forecasting task. This work contributes to the advancement of NLP by reframing qualitative time-series forecasting as a language problem.

Our main contributions are summarized as follows.

- A three-tier LLM framework tackling qualitative LE data's noise and MNAR missingness via imputation and feature selection.
- A novel zero-shot LLM selection method, outperforming numeric baselines on textual time-series.
- Evidence of LLMs' superiority for sparse, subjective sequences, advancing NLP's temporal scope.

2 Related Work

This research leverages LLMs for time-series forecasting, extending their NLP strengths to qualitative LE data in education. Transformer-based LLMs like TimeGPT (Garza et al., 2024) and PromptCast (Xue and Salim, 2024) verbalize numeric time-series for prediction, with data-centric approaches transforming sequences into text for pre-trained LMs (Jin et al., 2024) and model-centric methods fine-tuning LMs for temporal tasks (Zhou et al., 2023). Our model-centric approach fine-tunes LLMs for subjective LE sequences, diverging from numeric trends to target engagement attributes—a domain underexplored by existing LLM-based time-series models despite their sequential modeling prowess.

Student engagement forecasting in educational analytics often relies on cognitive (e.g., grades) or behavioral (e.g., clickstreams) data, using ML methods like LSTMs and Random Forests (Xu and Ouyang, 2022). Recent work incorporates NC factors—self-efficacy, motivation—from surveys (Fredricks, 2014), yet struggles with textual responses, temporal dynamics, and MNAR missingness prevalent in LE data (Sinatra et al., 2015). Unlike these numeric-focused efforts, our framework verbalizes weekly NC trajectories for LLM processing, forecasting binary engagement levels and bridging educational analytics with NLP's textual capabilities, addressing a gap in longitudinal engagement modeling.

Handling missing data and feature selection in LE sequences poses further challenges. Traditional imputation (e.g., MICE (van Buuren and Groothuis-Oudshoorn, 2011)) assumes MCAR/MAR, faltering with MNAR patterns (e.g., disengagement-driven skips) and LE's qualitative heterogeneity (Rubin, 1976). Generative models like GAIN (Yoon et al., 2018) impute numeric values but lack context for textual NC features, while standard feature selection (e.g., variance thresholding (Jain et al., 2000)) misses nuanced semantic relevance. Our three-tier framework—LLM-informed imputation (GPT-4o), zero-shot feature selection, and fine-tuned forecasting—outperforms these by capturing MNAR context and selecting predictive NC subsets, leveraging LLMs' reasoning for sparse, subjective data. See Appendix A.2 for a detailed discussion.

3 Three-Tier LLM Framework

This section details our three-tier NLP framework for forecasting weekly student engagement levels from qualitative LE data, designed to address the challenges of MNAR missingness and noise in the feature space. The framework, illustrated in Figure 1, consists of: (1) LLM-informed imputation to address MNAR gaps, (2) zero-shot feature selection via an ensemble of expert LLMs, and (3) fine-tuned forecasting with diverse LLM architectures. These tiers transform sparse, qualitative NC sequences into predictive models, evaluated against numeric baselines.

3.1 Dataset

We utilize a dataset from 96 first-year college students in introductory programming courses at a U.S. public university, collected over 15 weeks per semester across three semesters (Hayat et al., 2024a,b). The data captures 78-dimensional academic experiential trajectories across three modalities: 9-dimensional background data (e.g., demographics, socioeconomic status), 41-dimensional cognitive data (e.g., quiz scores, coding assignment grades), and 28-dimensional NC data (e.g., self-reported motivation, lecture engagement). Background data derives from an initial web survey, cognitive data from the course learning management system, and NC data from daily, context-adaptive questions via a privacy-preserving smartphone app, stored anonymously on cloud servers.

For forecasting, we focus on the NC data, comprising responses to 28 questions targeting behavioral, emotional, and cognitive engagement dimensions (e.g., "How much are you looking forward to your CS1 class lecture today?"). Due to high missingness—over 90% for 18 questions, with some entirely unanswered—we curated 10 key qualitative NC features with at least 35% response rates, detailed in Appendix A.1. These 10 features represent our curated set of key qualitative non-cognitive indicators of student engagement, chosen after addressing the issue of high missingness in the initial 28 features. Using a sliding window, we construct 4-week sequences to predict the subsequent week's engagement shift (e.g., weeks 1-4 predict week 5), yielding 960 trajectories (96 students × 10 predictions per semester). Each trajectory targets **four binary engagement outcomes**—Lecture Engagement Disposition, Academic Self-Efficacy, Performance Self-Evaluation, and Academic Iden-

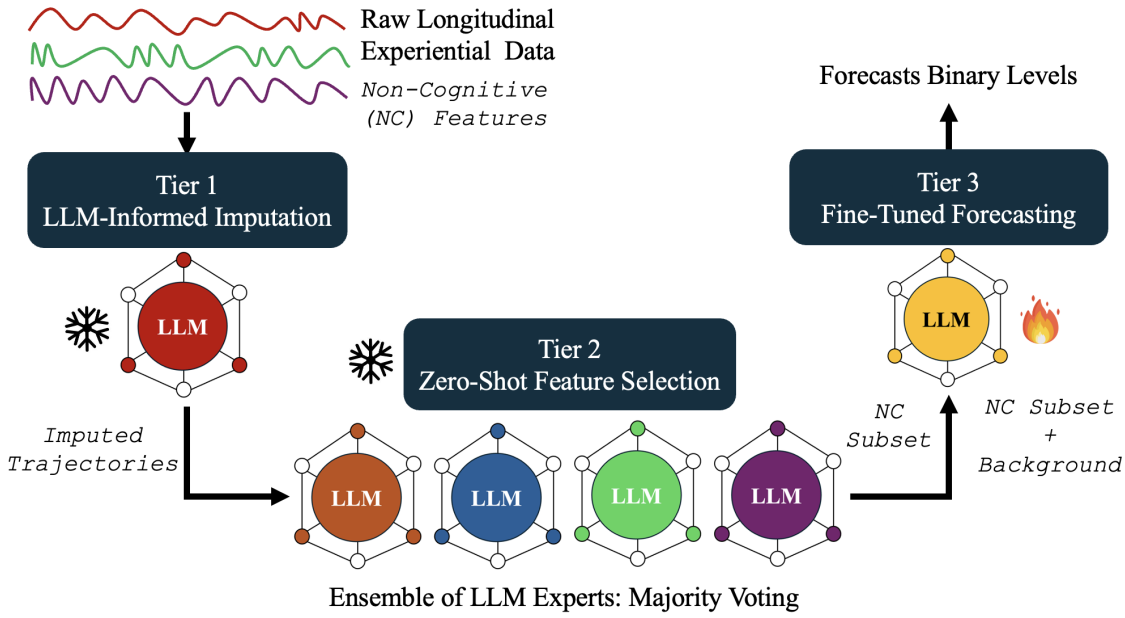

Figure 1: Three-tier LLM framework: (1) LLM-informed imputation fills MNAR gaps in LE trajectories, (2) zero-shot feature selection by expert LLMs curates NC subsets, and (3) fine-tuned LLMs forecast binary engagement levels, enhanced by background features.

tity and Value Perception—derived as composite scores from the 10 NC features.

3.2 LLM-Informed Imputation

The NC data exhibits significant missingness (e.g., 66% of responses missing in week 1, 37% of students skipping questions for over two weeks), often due to students skipping questions or uninstalling the app—patterns indicative of MNAR behavior (Rubin, 1976). Traditional imputation methods like LOCF (Liu, 2016) are unsuitable, as entire weekly response sets may be absent, leaving no prior values to propagate, and numeric imputation risks introducing bias by ignoring MNAR's semantic context (Little and Rubin, 2019). To address this critical issue of biased missingness, we employ GPT-4o (OpenAI, 2024) in a zero-shot manner to generate textual descriptors for missing responses (e.g., "The student skipped this question" or "No response due to app uninstallation"), preserving contextual meaning without forcing numeric assumptions.

For each missing response in our dataset, we construct a detailed prompt that includes surrounding NC data, such as responses from prior or subsequent weeks, and contextual metadata, such as question type and week number. This information serves as the dataset features information provided to GPT-4o, enabling it to infer a descriptor for the missing response. GPT-4o processes these prompts zero-shot—without task-specific training—leveraging its linguistic reasoning to infer descriptors that reflect MNAR dynamics (e.g., disengagement patterns). This approach enhances data quality by embedding semantic context into the 960 trajectories, enabling downstream feature selection and forecasting to exploit qualitative signals overlooked by statistical methods (Little and Rubin, 2019).

3.3 Zero-Shot Feature Selection

Our dataset's 10 curated NC features, reduced from an initial set of 28 due to extreme missingness, form a semantically rich yet sparse space requiring feature selection to optimize forecasting by reducing noise inherent in the qualitative feature space (Guyon and Elisseeff, 2003). Traditional methods—e.g., variance thresholding, correlation analysis (Jain et al., 2000), or attention-based deep learning (Ying et al., 2024)—rely on statistical distributions or labeled data, often missing qualitative, non-linear relationships in LE sequences. Instead, we propose a zero-shot feature selection method using **an ensemble of expert LLMs**: GPT-4o (OpenAI, 2024), Google Gemini (Team et al., 2024a), DeepSeek (DeepSeek-AI et al., 2025), and Mi-

crosoft Copilot (Copilot, 2024). This panel leverages each model's linguistic reasoning and world knowledge to identify predictive NC subsets for four engagement dimensions without accessing the data itself (Kojima et al., 2022).

Mathematical Formulation. Formally, let $\mathbf{X}$ denote the dataset of student responses, where each instance $\mathbf{X}^i = (X_1^i, \ldots, X_d^i)$ is a d-dimensional vector, and $d = 10$ represents the curated NC features (e.g., X_1^i: motivation, X_5^i: lecture enjoyment). For each engagement dimension k (e.g., $k =$ LED for Lecture Engagement Disposition), we define a candidate feature set $\mathbf{F} = \{X_1, \ldots, X_{10}\}$ and seek an optimal subset $\mathbf{S}_k \subseteq \mathbf{F}$ that maximizes predictive relevance for target Y_k. Unlike statistical methods that require data access, our ensemble operates zero-shot: given only semantic descriptions of the 10 features and target definitions, each LLM M_j independently produces a ranking R_j^k of features by inferred relevance. We aggregate these rankings via majority voting, where a feature X_i is included in $\mathbf{S}_k$ if selected by at least $\lceil J/2 \rceil$ models, where $J = 4$ is the ensemble size. This yields consensus-driven subsets $\mathbf{S}_k$ that capture semantic relationships across multiple expert perspectives.

Implementation via Unified Expert Prompt. To systematically guide the feature selection process across all four engagement dimensions, we employ a single comprehensive prompt that leverages psychological domain expertise. Each LLM in our ensemble receives the following structured prompt:

> *"You are an expert psychologist analyzing and predicting student engagement. Given a set of survey questions (e.g., **Q1**: How much are you looking forward to today's lecture?, **Q5**: How much did you enjoy today's lecture?, **Q18**: How confident are you in your programming skills?, . . .), identify the most predictive ones for forecasting students' engagement levels in the following domains for the upcoming week:"*
>
> 1. *Lecture Engagement Disposition*
> 2. *Academic Self-Efficacy*
> 3. *Performance Self-Evaluation*
> 4. *Academic Identity and Value Perception*

This unified prompting approach ensures consistency across the ensemble while allowing each LLM to apply its domain knowledge to identify dimension-specific feature subsets. The streamlined prompt structure enables each model to consider all available features and make informed selections based on psychological theory and semantic relationships between questions and target constructs.

The systematic application of this expert prompt yields tailored feature subsets: Lecture Engagement Disposition $\mathbf{S}_{\text{LED}} = \{Q1, Q5\}$, Academic Self-Efficacy $\mathbf{S}_{\text{ASE}} = \{Q18, Q19, Q20\}$, Performance Self-Evaluation $\mathbf{S}_{\text{PSE}} = \{Q21, Q22, Q23\}$, and Academic Identity and Value Perception $\mathbf{S}_{\text{AIVP}} = \{Q24, Q25\}$. By reasoning over semantic nuance (e.g., prioritizing "lecture enjoyment" over "general motivation" for LED), the ensemble captures contextual relationships statistical methods overlook. This data-agnostic, scalable approach leverages LLMs' prior knowledge, offering a novel alternative to traditional feature selection for qualitative time-series tasks (Kojima et al., 2022).

We acknowledge that our current implementation relies on closed-source LLMs (GPT-4o, Google Gemini, DeepSeek, and Microsoft Copilot), though the framework is adaptable to open-source alternatives such as Llama (Touvron et al., 2023) or Mistral (Mistral AI, 2024) for enhanced reproducibility.

3.4 Data Preprocessing

To generate binary labels for our 960 trajectories, we score NC responses on a scale capturing engagement intensity (e.g., for X_1: "I am really looking forward to it" = 1, "I am not planning to attend" = -1, "I am kind of looking forward" = 0.5, "I am not really looking forward" = -0.5). For each student and week, we compute a composite score per dimension by averaging the subset's scores selected for that dimension (e.g., for Lecture Engagement Disposition: $(X_1 + X_5)/2$, aggregating daily responses). For a 4-week sequence (e.g., weeks 1–4), we calculate the week 5 score; a positive shift ($Y_k = 1$) is assigned if the week 5 score exceeds the 4-week average, otherwise negative ($Y_k = 0$), yielding a positive-to-negative ratio between 60:40 and 70:30 across the four dimensions.

For **baseline models**, we convert NC responses into numeric features using these assigned scores (e.g., $X_1 = 1$ for "I am really looking forward to it"), preserving the 10-feature structure post-

selection (Section 3.3). Missing values, reflecting MNAR patterns, are imputed with zeros, forming 36-D vectors (10 features $\times$ 4 weeks) or 4 $\times$ 10-D sequences for model input. For **LLMs**, we verbalize these imputed 4-week sequences into natural language narratives, integrating GPT-4o descriptors from Section 3.2 to leverage text-processing strengths (Radford et al., 2019). For example, a sequence might read: "Week 1, Monday: Prior to the lecture, the student reported *I am not looking forward to it*; in the evening, they reflected: *I did not enjoy the lecture*. Week 2: [imputed] *skipped the question*". This transformation embeds daily responses (e.g., X_1: motivation, X_5: enjoyment) and imputed MNAR patterns, tailored to the four engagement dimensions, preparing data for fine-tuning with qualitative context intact.

3.5 Fine-Tuned Forecasting with LLM Architectures

We forecast binary engagement levels (positive vs. negative) over 10 weeks (weeks 5–14) using the 960 verbalized trajectories (Section 3.4). *Two different LLM architecture classes* are fine-tuned for binary classification of student engagement across key dimensions: decoder-only models (Gemma2 9B (Team et al., 2024b), Mixtral 8x7B (Jiang et al., 2024), Llama 7B (Touvron et al., 2023)) and encoder-only models (RoBERTa (Liu et al., 2021), DistilBERT-base-uncased (Sanh et al., 2020)).

Decoder-only models leverage autoregressive reasoning to model the narrative complexity and temporal dependencies of verbalized NC trajectories (e.g., 4-week sequences with MNAR-imputed text), potentially capturing nuanced shifts in qualitative LE data. Specifically, we include Gemma2 9B for its strong performance and efficiency, Mixtral 8x7B as a sparse mixture-of-experts model known for its high quality and fast inference, and Llama 7B as a widely adopted and well-studied foundational model.

Conversely, encoder-only models excel at bidirectional sequence encoding, optimizing discriminative power for sparse, noisy inputs by focusing on contextual feature interactions—critical for our 960 trajectories with varying missingness (35%–100%). We select RoBERTa for its robust pretraining and state-of-the-art results on various classification tasks, and DistilBERT-base-uncased as a computationally efficient yet effective transformer model, allowing us to explore the trade-off between model size and performance.

This dual selection tests architectural suitability: generative flexibility for sequential coherence vs. compact representation for classification efficacy.

Performance is evaluated using balanced accuracy and macro-F1, against numeric baselines: Random Forest (Breiman, 2001), Support Vector Machines (Hearst et al., 1998), 1D CNN (O'Shea and Nash, 2015), LSTM (Hochreiter and Schmidhuber, 1997), and Transformer (Vaswani et al., 2017). Baselines use scored responses (no verbalization).

4 Experiments and Results

We evaluate our three-tier LLM framework—LLM-informed imputation, zero-shot feature selection, and fine-tuned forecasting—against numeric baselines to demonstrate its effectiveness in forecasting student engagement from qualitative LE data. Key comparisons assess: (1) baseline machine and deep learning models with numeric NC subset features, (2) LLMs with verbalized NC subset features, and (3) LLMs with NC subset plus background features. Two ablation studies further explore feature quantity (subset vs. all NC features) and input modality (textual LLMs vs. numeric baselines with all features), validating LLMs' superiority and selection benefits for noisy, MNAR-impaired data.

4.1 Experimental Setup

The dataset comprises 960 trajectories (Section 3.1), split into 70% training (672 trajectories), 15% validation (144), and 15% testing (144), with positive-to-negative class ratios ranging from 60:40 to 70:30 across four dimensions: Lecture Engagement Disposition (LED), Academic Self-Efficacy (ASE), Performance Self-Efficacy (PSE), and Academic Identity and Value Perception (AIVP). Three configurations are tested: (1) numeric NC subset features (e.g., LED: Q1, Q5 post zero-shot selection, Section 3.3), (2) verbalized NC subset features, and (3) verbalized NC subset features plus 9 background features (e.g., demographics) appended as "Background: Female, Mechanical Engineering Major...".

Baseline models—Random Forest (RF, 100 trees), SVM (RBF kernel), 1D CNN, LSTM, and Transformer—are implemented via scikit-learn, trained on numeric NC subset features (e.g., LED: Q1, Q5 as scores, Section 3.4). The 1D CNN uses two convolutional layers with max-pooling, followed by fully connected and dropout layers.

Table 1: Baseline Performance Across Dimensions Using Numeric NC Subset Features

Model	LED		ASE		PSE		AIVP	
	B.Acc.	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**
Random Forest	54.5	53.5	46.0	44.5	**53.5**	**52.5**	44.0	41.0
SVM	52.0	48.0	50.0	41.0	51.5	47.0	50.0	39.0
1D CNN	49.0	49.0	48.5	46.0	42.5	39.5	50.0	**41.5**
Transformer	49.5	47.0	50.0	**50.0**	50.0	46.5	49.5	38.5
LSTM	**55.5**	**54.0**	**53.5**	48.0	47.5	45.0	**51.5**	40.5

The LSTM employs two 50-unit layers, the first returning sequences, with dropout. The Transformer features two MultiHeadAttention layers, feed-forward networks, and global average pooling, with dropout. These run on 8× NVIDIA A40 GPUs with a batch size of 32, learning rate of 0.001, 50 epochs, and AdamW optimizer (Loshchilov and Hutter, 2019). LLMs—decoder-only (Gemma2 9B (Team et al., 2024b), Mixtral 8x7B (Jiang et al., 2024), LLaMA 7B (Touvron et al., 2023)) and encoder-only (RoBERTa (Liu et al., 2021), DistilBERT (Sanh et al., 2020))—are fine-tuned via Hugging Face Transformers on the same GPUs, with a batch size of 8, learning rate of 1×10^{-5}, 20 epochs, and AdamW with weight decay 0.01. Given class imbalance, we report balanced accuracy (B.Acc.) and macro-F1 score.

4.2 Results and Analysis

4.2.1 Baseline Performance

Table 1 shows the performance of baseline models trained on numeric NC subset features (e.g., LED: Q1, Q5; Section 3.3) across four dimensions: LED, ASE, PSE, and AIVP. LSTM leads in three dimensions, with balanced accuracy (B.Acc.) of 55.5% (LED), 53.5% (ASE), and 51.5% (AIVP), and macro-F1 peaking at 54.0% (LED), leveraging its sequential modeling capability. Random Forest (RF) excels for PSE (53.5% B.Acc., 52.5% F1), surpassing LSTM through robust feature aggregation. SVM and Transformer achieve moderate results, with Transformer's best F1 at 50.0% (ASE), while 1D CNN consistently underperforms (e.g., 39.5% F1 for PSE). Across dimensions, baselines average 50.8% balanced accuracy and 46.9% macro-F1, struggling with sparsity and MNAR patterns. These models have demonstrated a tendency to deliver unreliable results, with a significant skew towards predicting outcomes predominantly in the positive class, which makes these models unreliable for these tasks.

4.2.2 LLM Performance with NC Subset Features

Table 2 presents the performance of fine-tuned LLMs using verbalized NC subset features (e.g., 2–3 features per dimension, Section 3.3) across all four dimensions: LED, ASE, PSE, and AIVP. RoBERTa consistently achieves the highest macro-F1 scores, ranging from 65.0% (LED) to 70.5% (ASE, AIVP), with balanced accuracy peaking at 69.0% (AIVP), surpassing the best baseline (LSTM, 54.0% F1 for LED) by 11%–17%. Encoder-only models outperform decoder-only counterparts, with DistilBERT close behind RoBERTa (e.g., 68.5% F1 for AIVP vs. 70.5%), while decoder-only models show variability: Llama excels for PSE (73.0% F1) but drops to 56.5% for LED, and Mixtral lags across dimensions (55.5%–63.0% F1). Gemma2 performs well for ASE (70.0% F1) but averages lower elsewhere. The mean balanced accuracy (64.2%) and macro-F1 (64.4%) of LLMs highlight their textual reasoning advantage over numeric baselines, supporting their baseline superiority.

4.2.3 LLM Performance with NC Subset + Background Features

Table 3 reports LLM performance when NC subset features are augmented with background data (e.g., demographics). RoBERTa again dominates, with balanced accuracy improving to 72.5%–77.5% and macro-F1 to 73.0%–77.5% across dimensions, a 3%–12% gain over NC-only results (e.g., LED F1: 65.0% to 77.5%). This boost peaks for LED (77.5% F1), affirming background data's contextual value. DistilBERT follows closely, with notable gains (e.g., LED F1: 64.5% to 75.0%), while decoder-only models improve but remain inconsistent: Llama reaches 74.5% F1 for LED but dips to 66.5% for PSE, Gemma2 holds steady (e.g., 70.0% F1 for AIVP), and Mixtral trails (61.0%–66.0% F1). The mean balanced accuracy rises to 69.0% and macro-F1 to 69.5%, with encoder-only models (RoBERTa: 74.5% mean F1, DistilBERT: 68.3%) outperforming decoder-only

Table 2: LLM Performance Across Dimensions Using NC Subset Features Only

Model	LED		ASE		PSE		AIVP	
	B.Acc.	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**
Gemma2 9B	62.0	61.0	**72.0**	70.0	65.5	66.5	65.5	66.5
Mixtral 8x7B	62.0	61.5	63.5	63.0	55.5	55.5	59.0	59.0
Llama 7B	59.5	56.5	62.0	61.5	**73.0**	**73.0**	59.5	59.0
DistilBERT	65.0	64.0	63.5	67.0	67.5	67.0	67.5	68.5
RoBERTa	**65.0**	**65.0**	66.5	**70.5**	68.0	69.5	**69.0**	**70.5**

Table 3: LLM Performance Across Dimensions Using NC Subset and Background Features

Model	LED		ASE		PSE		AIVP	
	B.Acc.	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**
Gemma2 9B	71.5	72.0	69.0	69.0	65.5	64.5	69.0	70.0
Mixtral 8x7B	60.0	61.5	66.0	66.0	61.5	61.5	61.0	61.0
Llama 7B	72.5	74.5	68.0	69.0	66.5	66.5	66.5	66.5
DistilBERT	74.5	75.0	65.5	66.0	70.5	68.0	65.0	64.5
RoBERTa	**77.5**	**77.5**	**73.5**	**73.0**	**74.0**	**73.5**	**72.5**	**74.0**

Table 4: LLM Performance Across Dimensions Using All NC and Background Features

Model	LED		ASE		PSE		AIVP	
	B.Acc.	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**
Gemma2 9B	58.50	52.50	62.50	65.50	59.50	58.50	66.00	65.50
Mixtral 8x7B	55.50	56.50	59.00	63.00	58.00	56.50	52.00	53.50
Llama 7B	59.50	56.50	62.50	62.50	59.00	59.00	60.00	63.50
DistilBERT	62.00	62.50	64.00	63.50	60.50	62.50	65.50	64.00
RoBERTa	**66.50**	**65.00**	**64.50**	**65.00**	**63.00**	**62.50**	**64.50**	**64.00**

(Llama: 69.1%, Gemma2: 67.9%, Mixtral: 62.5%) by 5%–12%. Compared to baselines (max 54.0% F1), NC+background LLMs extend the gap to 19%–23%.

4.3 Ablation Study

We conduct two ablation studies to evaluate our LLM-based approach for forecasting student engagement levels across four dimensions.

Evaluating Feature Quantity: Subset vs. All NC Features. We compare LLMs fine-tuned on a zero-shot selected subset of NC features plus background features (Table 3) against those using all 10 NC features plus background features (Table 4). In the all-features case, RoBERTa achieves macro-F1 scores of 62.5%–65.0% and balanced accuracy (B.Acc.) of 63.0%–66.5%, markedly lower than the subset case's 73.0%–77.5% F1 and 72.5%–77.5% B.Acc. Dimension-specific F1 losses range from 8.0% (ASE) to 12.5% (LED), indicating that all 10 NC features introduce noise, weakening the signal distilled by expert LLM selection (Section 3.3). Encoder-only models (RoBERTa, DistilBERT) consistently outperform decoder-only variants (Gemma2 9B, Mixtral 8x7B, LLaMA 7B) across both configurations, though the gap narrows with all features—e.g., RoBERTa's LED F1 lead over LLaMA 7B shrinks from 18.0% (subset) to 8.5%—suggesting noise impacts decoder-only models less severely.

Assessing Input Modality: Textual LLMs vs. Numeric Baselines with All Features. We train baseline models—Random Forest (RF, 100 trees (Breiman, 2001)), Support Vector Machine (SVM), 1D CNN, Transformer, and LSTM—on numeric LE data with all 10 NC features (converted to scores, forming 960 × 36-D vectors, Section 3.4) and fine-tune RoBERTa (Liu et al., 2021), our top performer with subset features, on textual all NC features (verbalized responses). Table 5 reports results across 960 trajectories for four dimensions: LED, ASE, PSE, and AIVP. RoBERTa consistently outperforms numeric baselines in balanced accuracy and macro-F1 across most dimensions, leveraging textual reasoning to capture qualitative nuances and MNAR-impaired patterns that numeric models struggle to model. Notably, 1D CNN excels for ASE, suggesting some sequential patterns in numeric data align with convolutional strengths, yet RoBERTa's broader superiority—particularly for LED, PSE, and AIVP—underscores LLMs' advantage in processing raw verbalized sequences. Baselines like SVM and LSTM exhibit variability, often skewed by noise or positive-class bias, while Transformer and RF show moderate consistency but lack the discriminative power of textual LLMs. This complements the first ablation study (subset vs. all NC features), affirming that while subset selection enhances performance, even with all features, LLMs' textual modality outstrips numeric approaches for sparse, qualitative LE data.

Table 5: Performance of Numeric Baselines and Textual RoBERTa with All NC Features Across Dimensions

Model	LED		ASE		PSE		AIVP	
	B.Acc.	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**	**B.Acc.**	**F1**
Random Forest	52.5	49.5	50.5	45.5	49.0	46.5	48.0	41.0
SVM	50.5	45.5	50.0	40.5	44.5	40.0	50.0	38.0
1D CNN	52.0	51.5	**62.0**	**61.5**	48.0	47.5	44.5	43.0
Transformer	48.5	47.5	45.5	45.0	51.5	51.5	53.5	53.0
LSTM	44.0	43.5	47.5	47.5	47.0	47.0	50.0	47.5
RoBERTa	**61.0**	**60.5**	59.0	65.0	**64.0**	**62.5**	**62.5**	**62.0**

5 Conclusion

Our findings provide compelling evidence for the efficacy of our three-tier LLM framework in forecasting student engagement from qualitative LE data. We highlight three key insights. First, LLMs consistently outperformed traditional numeric baselines across all engagement dimensions, even when both were trained on the same selected non-cognitive feature subsets. This superiority underscores the inherent capability of LLMs to process and understand the nuanced information present in verbalized student responses, effectively capturing contextual patterns missed by numeric conversions and sequential models, particularly in the presence of MNAR missingness and data sparsity. Notably, this advantage persisted even when all available non-cognitive features were used, further emphasizing the limitations of traditional machine learning approaches for this type of qualitative time-series data.

Second, our analysis revealed a significant performance difference between LLM architectures. Encoder-only models, such as RoBERTa and DistilBERT, demonstrated a clear advantage over decoder-only models across various configurations. This suggests that their strength in creating robust representations from sparse textual sequences makes them particularly well-suited for the binary classification task of engagement forecasting. While decoder-only models showed occasional strong performance on specific dimensions, their overall variability indicates that their generative focus might be less optimal for the discriminative demands of this task. The consistent outperformance of encoder-only models, even with increased data complexity, highlights their robustness for this application.

Third, integrating background data significantly boosted LLM performance, particularly for specific engagement dimensions, emphasizing the importance of context. Furthermore, the synergy between our LLM-driven feature selection and forecasting tiers was validated by the enhanced performance achieved with selected feature subsets.

In conclusion, this work demonstrates the transformative potential of our three-tier LLM framework for analyzing complex, qualitative LE data in educational settings. By effectively addressing challenges such as MNAR missingness and noisy feature spaces, our approach offers a significant advancement over traditional numeric methods, paving the way for richer and more insightful analyses of student engagement and potentially other subjective, time-series datasets. **However, responsible deployment of such frameworks requires careful consideration of their limitations and ethical implications.**

6 Limitations

Our study acknowledges several important limitations. **Dataset scale and diversity**: Our analysis is based on data from 96 first-year STEM students at a single U.S. university, resulting in 960 trajectories. This relatively small and homogeneous sample limits generalizability to broader student populations, diverse educational contexts, or different demographic groups. **Validation constraints**: Our LLM-informed imputation method has not undergone human validation to verify the accuracy of generated missing value descriptors, affecting confidence in semantic quality and downstream forecasting performance. **Baseline limitations**: Our evaluation focuses on traditional machine learning and basic deep learning models, but does not benchmark against state-of-the-art multimodal or recent transformer-based time-series forecasting models. **Theoretical justification**: While empirical results demonstrate encoder-only LLMs' superior performance over decoder-only models, we provide limited theoretical explanation for this architectural advantage. **Dependency on proprietary models**: Our framework relies on closed-source LLMs (GPT-4o, Gemini, DeepSeek, Copilot), which may limit reproducibility and accessibility.

7 Ethical Considerations

7.1 LLM Biases and Educational Harms

Foundation models encode systemic biases from pretraining data (Bommasani et al., 2021), which can be amplified when fine-tuned on small educational datasets. LLMs characterized as "stochastic parrots" (Bender et al., 2021) exhibit stereotypical biases across gender, race, profession, and religion (Nadeem et al., 2021; Gallegos et al., 2024), with documented religious bias analogizing "Muslim" to "terrorist" in 23% of cases (Abid et al., 2021). A comprehensive risk taxonomy identifies discrimination, hate speech, and human-computer interaction harms as primary concerns (Weidinger et al., 2022).

Our framework's LLM-informed imputation and feature selection may inadvertently reflect these biases, potentially misrepresenting underrepresented student voices or reinforcing stereotypical engagement assumptions. Algorithmic bias in education disproportionately affects students based on race/ethnicity, gender, nationality, socioeconomic status, and disability (Baker and Hawn, 2022). Automated engagement predictions risk reinforcing inequalities through biased classifications that systematically disadvantage certain groups, as foundation model defects are inherited downstream (Bommasani et al., 2021).

7.2 Potential Harms and Mitigation

Self-fulfilling prophecies: Predictions may influence educator expectations, creating scenarios where students labeled "disengaged" receive reduced support. Automated decision-making risks "reducing a human being to a percentage," undermining student dignity (Binns et al., 2018). **Student autonomy**: Engagement monitoring may create surveillance environments compromising authentic self-expression and altering social dynamics the technology purports to measure (Weidinger et al., 2022). **Resource allocation**: Binary predictions could lead to misallocation if false positives/negatives disproportionately affect vulnerable populations (Corbett-Davies et al., 2017).

Privacy considerations: Our dataset involves sensitive student information including academic performance and personal reflections. While committing to full anonymization, evolving LLM capabilities may create unforeseen privacy risks not understood at consent time.

Mitigation strategies: We propose safeguards informed by responsible AI principles (Weidinger et al., 2022; Bommasani et al., 2021): (1) Regular bias auditing across demographic subgroups; (2) Human-in-the-loop validation requiring educator oversight before interventions (Binns et al., 2018); (3) Transparent communication about data use; (4) Supportive-only intervention guidelines; (5) Continuous monitoring of deployment outcomes. Our framework should augment, not replace, human educational judgment, emphasizing fairness, accountability, and transparency in high-stakes educational applications (Binns et al., 2018; Weidinger et al., 2022; Baker and Hawn, 2022).

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

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

Ryan S. Baker and Andrew Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32:1052–1092.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy,

Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*, 5th edition. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Leo Breiman. 2001. Random Forests. *Mach. Learn.*, 45(1):5?32.

Microsoft Copilot. 2024. Generated content. Online. Accessed on January 13, 2025.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2024. Timegpt-1. *Preprint*, arXiv:2310.03589.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null):1157–1182.

Ahatsham Hayat, Bilal Khan, and Mohammad Hasan. 2024a. Improving transfer learning for early forecasting of academic performance by contextualizing language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 137–148, Mexico City, Mexico. Association for Computational Linguistics.

Ahatsham Hayat, Bilal Khan, and Mohammad Rashedul Hasan. 2024b. Leveraging language models for analyzing longitudinal experiential data in education. *arXiv:2503.21617*.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

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

A.K. Jain, R.P.W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

David A. Kolb. 1984. *Experiential Learning: Experience as the Source of Learning and Development*. Prentice-Hall, Englewood Cliffs, NJ.

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.

R. J. A. Little and D. B. Rubin. 2019. *Statistical Analysis with Missing Data*. John Wiley & Sons.

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.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings*, page 471–484, Berlin, Heidelberg. Springer-Verlag.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Mistral AI. 2024. Announcing mistral 7b instruct v0.3. `https://mistral.ai/news/announcing-mistral-7b/`. Accessed: 2025-05-20.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-09-02.

Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458.

M. Palmer, M. Larkin, R. de Visser, and G. Fadden. 2010. Developing an interpretative phenomenological approach to focus group data. *Qualitative Research in Psychology*, 7(2):99–121.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

D. B. Rubin. 1976. Inference and missing data. *Biometrika*, 63:581–592.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

J. L. Schafer. 1997. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, UK.

Gale M. Sinatra, Benjamin C. Heddy, and Doug Lombardi. 2015. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak

Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

S. van Buuren and K. Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

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.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. 2022. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. In *Advances in Neural Information Processing Systems*, volume 35, pages 24655–24692. Curran Associates, Inc.

Hao Xue and Flora D. Salim. 2024. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Trans. on Knowl. and Data Eng.*, 36(11):6851–6864.

Wangyang Ying, Dongjie Wang, Haifeng Chen, and Yanjie Fu. 2024. Feature selection as deep sequential generative learning. *Preprint*, arXiv:2403.03838.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5689–5698. PMLR.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One fits all: power general time series analysis by pretrained lm. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.