ConText-LE: Cross-Distribution Generalization for Longitudinal Experiential Data via Narrative-Based LLM Representations

Anonymous ACL submission

Abstract

Longitudinal experiential data offers rich insights into dynamic human states, yet building models that generalize across diverse contexts remains challenging. This paper addresses how to best represent multi-modal longitudinal experiential data as text and formulate prediction tasks to maximize large language model (LLM) cross-distribution generalization. We propose ConText-LE, a framework grounded in linguistic and cognitive theories of contextual meaning-making, which systematically investigates text representation strategies and output formulations for robust behavioral forecasting. Our novel Meta-Narrative representation synthesizes complex temporal patterns into semantically rich narratives, while Prospective Narrative Generation reframes prediction as a generative task aligned with LLMs' inherent contextual understanding capabilities. Through comprehensive experiments on three diverse longitudinal datasets, we address the critical but underexplored challenge of cross-distribution generalization in mental health and educational behavior forecasting. We demonstrate that combining Meta-Narrative input with Prospective Narrative Generation significantly outperforms existing LLM-based approaches, achieving up to 12.28% improvement in out-of-distribution accuracy and up to 11.99% improvement in F1 scores over binary classification methods. Bidirectional evaluation and architectural ablation studies confirm the robustness of our approach, establishing ConText-LE as an effective framework for developing reliable behavioral forecasting systems across temporal and contextual shifts.

1 Introduction

Longitudinal experiential (LE) data—collected through Experience Sampling Methods (ESM) (Larson and Csikszentmihalyi, 1983), Ecological Momentary Assessment (EMA) (Stone and Shiffman, 1994; Shiffman et al., 2008), and passive sensing (Mohr et al., 2017; Kumar et al., 2015)—offers unprecedented opportunities to understand and predict dynamic human states in real-world contexts. By capturing both subjective reports (e.g., mood, stress) and objective measurements (e.g., activity, sleep patterns), LE data holds immense promise for personalized interventions in mental health (Xu et al., 2021a; Mohr et al., 2021) and education (Wang et al., 2014).

Despite this potential, a fundamental challenge remains largely unaddressed: **cross-distribution generalization**. Models trained on data from one cohort, time period, or context often exhibit dramatic performance degradation when applied to different populations or temporal periods (Xu et al., 2023a,b). This generalization failure represents a critical barrier to real-world deployment, as evidenced by the limited success of existing approaches when evaluated across distribution shifts. For instance, traditional machine learning approaches on the GLOBEM dataset achieve only 52.80% $\pm$ 0.024 out-of-distribution accuracy (Xu et al., 2023b), barely exceeding random chance.

We hypothesize that this generalization challenge stems from the *inherently contextual and situated nature of LE data*. Unlike traditional time series (Zhong et al., 2025), LE data carries implicit contextual meaning where the significance of behavioral patterns depends heavily on individual circumstances and broader social contexts. Consider a university student showing decreased activity and increased sleep during final exams—patterns that might indicate depression in other contexts but represent adaptive responses to academic stress in this specific situation.

Traditional machine learning approaches (Xu et al., 2019; Saeb et al., 2015; Wang et al., 2018) treat behavioral features as context-independent variables with fixed meanings. This limitation parallels early word embedding models that treated words as static vectors, before contextualized representations revolutionized NLP (Devlin et al., 2019; Peters et al., 2018). We propose that large language models (LLMs), with their pre-trained understanding of human behavior and contextual reasoning (Brown et al., 2020; Bommasani et al., 2022), offer unique capabilities for interpreting LE data within appropriate contexts.

However, existing LLM applications to LE data (Kim et al., 2024; Hayat et al., 2024a; Thach et al., 2025) have not systematically investigated cross-distribution generalization. They primarily employ simple text encodings (e.g., structured value lists, statistical summaries) paired with binary classification, overlooking how representation strategies and output formulations impact generalization performance. In our cross-distribution evaluation, these approaches show substantial performance drops, highlighting critical gaps in leveraging LLMs for robust behavioral modeling.

ConText-LE Framework: We introduce ConText-LE, a novel framework for generalizable LLM-based LE data modeling that systematically investigates the impact of textual representations and output formulations on cross-distribution performance. ConText-LE explores four distinct input representations:

- Three existing approaches: Complete Sequence (Hayat et al., 2024a), Statistical Summary (Thach et al., 2025), and Natural Language String (Kim et al., 2024)
- Our novel **Meta-Narrative**: High-level interpretative narratives that synthesize complex temporal patterns into semantically rich, contextually grounded summaries emphasizing feature relationships and potential real-world interpretations

We also compare two output formulations: traditional Binary Classification versus our proposed **Prospective Narrative Generation**, which reframes prediction as generating descriptive narratives about future states. This generative approach better aligns with LLMs' inherent capabilities and allows for more nuanced expression of contextual predictions.

Through comprehensive experiments on three diverse datasets (GLOBEM (Xu et al., 2023a), LifeSnaps (Yfantidou et al., 2022), and MFAFY (Hayat et al., 2024a,b; Thach et al., 2025)) focusing specifically on cross-distribution generalization—an underexplored but critical challenge—we demonstrate that combining Meta-Narrative input with Prospective Narrative Generation achieves superior performance. Our approach improves out-of-distribution accuracy by up to 12.28% and F1 scores by up to 11.99% compared to binary classification, establishing new benchmarks for robust behavioral forecasting across temporal and contextual shifts.

Our main contributions include:

- The **ConText-LE** framework for systematic investigation of textual representations and output formulations in LLM-based LE data modeling, addressing the critical challenge of cross-distribution generalization.
- **Meta-Narrative representation**, a novel two-stage technique that synthesizes complex temporal patterns into semantically rich narratives, and **Prospective Narrative Generation**, which reframes prediction as a generative task aligned with LLMs' contextual reasoning capabilities.
- Comprehensive empirical evaluation demonstrating substantial improvements (up to 12.28% accuracy, 11.99% F1) over existing approaches across three diverse datasets, establishing the first systematic benchmarks for cross-distribution behavioral forecasting.
- Architectural ablation studies revealing the critical importance of instruction tuning and context length for behavioral pattern interpretation, providing practical guidance for LLM selection in sensitive applications.

2 Related Work

Modeling LE Data: Longitudinal experiential data has been modeled using various traditional ML and deep learning approaches for healthcare (Wang et al., 2018; Xu et al., 2021a; Nemati et al., 2022) and education (Wang et al., 2016; Li et al., 2020). These methods often struggle with generalizability across domain shifts (Xu et al., 2023b) and inadequately handle missing data (Xu et al., 2021a; Arnold and Pistilli, 2012). Recent work has begun exploring LLMs for LE data forecasting (Kim et al., 2024; Hayat et al., 2024a; Thach et al., 2025), but primarily focuses on within-dataset evaluation rather than cross-distribution generalization.

NLP Foundations: The evolution from static word embeddings (Mikolov et al., 2013) to contextualized representations (Devlin et al., 2019) has revolutionized NLP by capturing how meaning changes with context. Recent advances in prompting strategies (Wei et al., 2023; Kojima et al., 2023) have enhanced LLMs' reasoning capabilities. Our work builds on these developments by treating multi-dimensional LE data as complex semantic structures requiring contextual interpretation, while leveraging findings that generative formulations often enable more effective reasoning than discriminative approaches.

Cross-Modal Applications: Recent work has explored adapting structured data for LLM processing through serialization or textual descriptions (Sun et al., 2023; Jin et al., 2023), with applications to human-centric data (Kim et al., 2024). Most approaches use simple encoding strategies, while our work investigates semantically rich narrative representations that better align with findings on how LLMs process contextual relationships (Wang et al., 2022a; Shwartz et al., 2020). A detailed review of related work is given in Appendix A.10.

3 The ConText-LE Framework

ConText-LE is a systematic framework for leveraging LLMs' contextual understanding capabilities to achieve robust cross-distribution generalization in LE data. Figure 1 illustrates the overall architecture, highlighting the interplay between textual representation strategies and output formulations.

3.1 Problem Formulation

Given LE data collected from N individuals over K weeks, we define feature vectors $\mathbf{x}_{i,j} \in \mathbb{R}^d$ for individual i at time step j, where d represents the dimensionality of multi-modal features (e.g., activity, sleep, mood, social interactions). Using a sliding window approach, we segment data into overlapping k-week sequences.

For cross-distribution generalization, we partition data into training period T and testing period T', where T' represents a different temporal period, cohort, or contextual setting. The model receives textual representation $X^{\text{text}}_{i,s:s+k-1}$ of each k-week sequence and predicts either a binary label $y^{\text{binary}}_{i,s+k} \in \{0, 1\}$ or narrative forecast $y^{\text{narrative}}_{i,s+k}$ for week $s + k$.

The core challenge lies in achieving robust performance when $P(X, Y|T) \neq P(X, Y|T')$, where distribution shifts may involve temporal changes (e.g., different academic semesters), demographic variations (e.g., different student cohorts), or contextual differences (e.g., pre/post-pandemic periods). Formal details are in Appendix A.1.

3.2 Textual Representation Strategies

ConText-LE investigates four distinct approaches for transforming raw LE data into textual inputs, each designed to capture different aspects of temporal and contextual information.

Baseline Representations We implement three existing approaches from prior work:

Complete Sequence (Hayat et al., 2024a): Direct verbalization of detailed temporal sequences. Example: *"Monday Jan 5: steps=8,245, heart_rate=72bpm, sleep=7.2hrs, mood=3/5. Tuesday Jan 6: steps=6,891, heart_rate=68bpm..."*

Statistical Summary (Thach et al., 2025): Aggregate statistics for each feature over the k-week period. Example: *"Steps: mean=7,834, std=2,451, min=1,023, max=15,672. Sleep: mean=7.1hrs, std=1.2hrs..."*

Natural Language String (Kim et al., 2024): Structured listing of feature values over time. Example: *"Steps: [8245, 6891, NaN, 9156, ...]; Sleep: [7.2, 6.8, NaN, 8.1, ...]; Mood: [3, 4, NaN, 2, ...]"*

Meta-Narrative Representation (Novel) Our proposed Meta-Narrative approach generates high-level interpretative narratives that synthesize complex temporal patterns into semantically rich, contextually grounded summaries. This representation is motivated by frame semantics theory (Fillmore, 2006), which suggests that meaning emerges from situating experiences within appropriate interpretive frameworks.

The Meta-Narrative is generated through a novel two-stage prompting process using GPT-4o (OpenAI, 2024):

Stage 1 - Feature Pattern Analysis: Identifies significant patterns in each behavioral dimension using statistical analysis and trend detection.

Stage 2 - Contextual Narrative Synthesis: Integrates individual patterns into a coherent narrative emphasizing inter-feature relationships, potential contextual interpretations, and global behavioral themes.

Example Meta-Narrative: *"This university student demonstrated consistent baseline activity patterns during the first three weeks, averaging 8,000 daily steps with regular 7-hour sleep cycles. However, week 4 marked a significant behavioral shift coinciding with the final examination period: activity decreased by 43% while sleep duration increased to over 9 hours nightly. Social interactions declined substantially from 8 to 2 weekly*

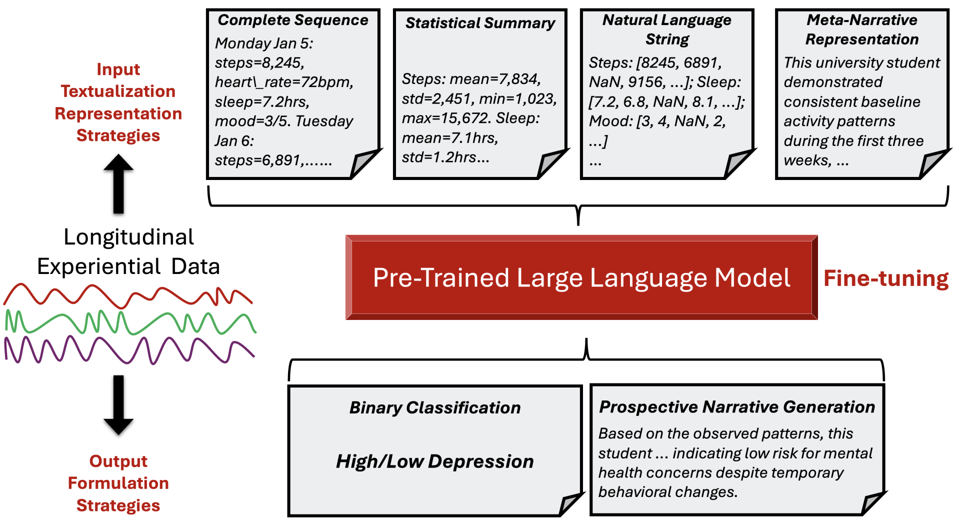

Figure 1: ConText-LE Framework Overview. The framework transforms multi-modal LE data through four representation strategies, processes them with fine-tuned LLMs using two output formulations, and evaluates cross-distribution generalization performance.

events. Despite these changes, self-reported mood remained stable at 'tired but OK,' suggesting adaptive rather than pathological responses to academic stress."

This approach transforms multi-dimensional time series into contextually rich narratives that better leverage LLMs' pre-trained understanding of human behavior patterns and situational interpretations. Prompt details are in Appendix A.4.

3.3 Output Formulations

ConText-LE compares two distinct approaches to formulating the prediction task, hypothesizing that generative formulations better align with LLMs' capabilities for contextual reasoning. A detailed description of these two formulations is provided in Appendix A.3.

Binary Classification The standard approach fine-tunes the LLM with a classification head to directly predict binary labels (e.g., low/high depression risk, academic engagement levels). This formulation treats prediction as a discriminative task requiring the model to compress complex behavioral patterns into a single binary decision.

Prospective Narrative Generation (Novel) Our proposed approach reframes prediction as a generative task where the LLM produces descriptive narratives about anticipated future states. This formulation is inspired by cognitive research on episodic future thinking (Schacter et al., 2008), where humans naturally predict future states through narrative construction rather than binary classification.

During training, target narratives $y_{i,s+k}^{\text{narrative}}$ are generated using GPT-4o to create coherent descriptions of future states that align with ground truth labels. During inference, the fine-tuned model generates prospective narratives from which binary predictions can be extracted if needed for evaluation.

Example target narrative: *"Based on the observed patterns, this student will likely experience continued academic stress in the upcoming week. Sleep patterns may remain elevated as exam preparation intensifies, while physical activity could decrease further. Social interactions will remain minimal, focused on study groups. Mood stability suggests effective coping mechanisms, indicating low risk for mental health concerns despite temporary behavioral changes."*

3.4 Model Architecture and Training

Base Model Selection We utilize Llama 3.1 8B Instruct (Grattafiori et al., 2024) as our foundation model, selected for its strong performance on language understanding tasks while maintaining computational efficiency suitable for extensive cross-distribution experiments.

Parameter-Efficient Fine-tuning Both output formulations employ Low-Rank Adaptation (LoRA) (Hu et al., 2021) for parameter-efficient fine-tuning. This approach adapts the model to LE data while preserving the pre-trained contextual knowledge crucial for generalization. LoRA enables efficient adaptation while maintaining most parameters frozen, reducing computational requirements and overfitting risks.

Training Strategy Models are trained separately for each textual representation and output formulation combination. For Prospective Narrative Generation, we employ teacher forcing during training with cross-entropy loss on generated tokens. Binary Classification uses standard cross-entropy loss on predicted labels. This systematic approach enables fair comparison across all framework components.

3.5 Evaluation Framework

Cross-Distribution Protocol Design Our evaluation protocol specifically targets cross-distribution generalization scenarios. We partition data into distinct temporal periods T (training) and T' (testing), ensuring no individual appears in both periods to prevent data leakage. This temporal splitting simulates realistic deployment scenarios where models must generalize to future time periods or different populations.

Evaluation Metrics We report standard binary classification metrics: accuracy, precision, recall, and F1-score, with primary focus on out-of-distribution performance. For narrative outputs, binary forecasts are extracted using GPT-4o with carefully designed prompts that maintain consistency across evaluations.

Baseline Establishment Strategy Given limited prior work on cross-distribution LE data generalization, we establish comprehensive baselines by re-implementing existing LLM approaches and adapting them for cross-distribution evaluation. Complete implementation details are provided in the experimental section.

4 Experiments and Results

4.1 Experimental Design

Datasets and Distribution Shifts We evaluate on three diverse LE datasets representing different types of distribution shifts:

GLOBEM (Xu et al., 2023a): Mental health prediction across 661 participants over 4 years. Cross-temporal shift: Years 1-2 (n=344, 2226 LE sequences) $rightarrow$ Years 3-4 (n=317, 2023 LE sequences). Features include activity, sleep, communication patterns, and mood assessments. Target: depression risk prediction.

LifeSnaps (Yfantidou et al., 2022): Anxiety prediction across 39 participants over 4 months. Cross-temporal shift: First 2 months (n=26, 112 LE sequences) $rightarrow$ Last 2 months (n=13, 64 LE sequences). Features include physiological signals, activity patterns, and self-reports. Target: anxiety episode prediction.

MFAFY (Hayat et al., 2024a): Academic engagement prediction across 96 participants over 2 years. Cross-temporal shift: Year 1 (2 semesters) (n=61, 610 LE sequences) $rightarrow$ Year 2 (1 semester) (n=35, 350 LE sequences). Features are qualitative self-reports of study behaviors and emotional states. Target: academic engagement level.

These datasets provide diverse modalities (structured sensors, physiological signals, unstructured text), scales (39-661 participants), and shift types (cohort changes, temporal dynamics, academic contexts), enabling robust evaluation of generalization capabilities. Detailed dataset information is in Appendix A.9.

4.2 Implementation and Evaluation Protocol

Implementation Details All experiments use Llama 3.1 8B Instruct (Grattafiori et al., 2024) with LoRA fine-tuning. Models are trained separately for each textual representation and output formulation combination. Textual representations and extractions use GPT-4o (OpenAI, 2024). Complete implementation details are in Appendix A.8.

Evaluation Metrics and Protocol We report accuracy, precision, recall, and F1-score, with primary focus on out-of-distribution performance. Data is partitioned into distinct temporal periods T (training: 85% train, 15% validation) and T' (testing: 100% OOD test), ensuring no individual appears in both periods. For narrative outputs, binary forecasts are extracted using GPT-4o with structured prompts.

Baseline Establishment We establish comprehensive baselines within the LLM framework by re-implementing three established textualization methods: Complete Sequence (Hayat et al., 2024a), Statistical Summary Encoding (Thach et al., 2025), and Natural Language String Encoding (Kim et al., 2024). For GLOBEM, we compare against the published cross-distribution baseline of 52.80% accuracy (Xu et al., 2023b).

4.3 Main Results: Cross-Distribution Performance

Table 1 presents comprehensive results across all datasets and configurations, revealing consistent patterns supporting ConText-LE's effectiveness.

Key Performance Patterns **Consistent Meta-Narrative Superiority**: Across all datasets and output formulations, Meta-Narrative achieves the

Table 1: Cross-distribution generalization results ($T \rightarrow T'$) across all datasets. Bold indicates best performance for each dataset, output formulation, and metric category.

Dataset	Shift	Input Strategy	In-Distribution (ID) Test				Out-of-Distribution (OOD) Test			
			Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
GLOBEM	Years 1&2 → Years 3&4	*Output Formulation: Binary Classification*								
		Complete Sequence	66.82	68.52	64.91	66.67	51.16	53.09	55.40	54.22
		Statistical Summary	63.68	64.81	61.95	63.35	51.11	53.08	54.73	53.89
		Natural Language String	67.26	70.00	65.81	67.84	52.64	53.54	56.95	55.19
		Meta-Narrative (ours)	**69.51**	**73.33**	**65.81**	**69.37**	**55.12**	**55.81**	**59.36**	**57.53**
		Output Formulation: Prospective Narrative Generation								
		Complete Sequence	69.96	71.56	68.42	69.96	65.94	67.95	68.52	68.23
		Statistical Summary	69.51	72.22	67.24	69.65	62.43	65.97	63.57	64.75
		Natural Language String	70.05	71.30	69.37	70.32	66.44	67.92	69.09	68.50
		Meta-Narrative (ours)	**73.99**	**75.93**	**71.93**	**73.87**	**67.40**	**68.81**	**70.00**	**69.40**
LifeSnaps	First 2 Months → Last 2 Months	*Output Formulation: Binary Classification*								
		Complete Sequence	58.82	62.50	55.56	58.82	51.56	44.12	55.56	49.18
		Statistical Summary	**82.35**	83.33	**90.91**	**86.96**	34.38	29.41	35.71	32.26
		Natural Language String	64.71	66.67	80.00	72.73	45.31	37.14	50.00	42.62
		Meta-Narrative (ours)	**82.35**	**90.00**	81.82	85.71	**59.38**	**53.12**	**60.71**	**56.67**
		Output Formulation: Prospective Narrative Generation								
		Complete Sequence	58.82	77.78	58.33	66.67	54.84	50.00	57.14	53.33
		Statistical Summary	47.06	40.00	57.14	47.06	46.88	36.67	42.31	39.29
		Natural Language String	**70.59**	**80.00**	**72.72**	**76.19**	62.50	52.94	69.23	60.00
		Meta-Narrative (ours)	64.71	77.78	63.64	70.00	**67.19**	**63.89**	**74.19**	**68.66**
MFAFY	Year 1 → Year 2	*Output Formulation: Binary Classification*								
		Complete Sequence	57.38	60.00	63.64	61.76	54.86	56.08	58.56	57.30
		Statistical Summary	45.90	34.48	41.67	37.74	48.86	49.18	51.14	50.14
		Natural Language String	57.38	58.33	65.62	61.76	59.83	47.52	50.00	48.73
		Meta-Narrative (ours)	**65.57**	**62.86**	**73.33**	**67.69**	**60.86**	**64.47**	**65.46**	**64.96**
		Output Formulation: Prospective Narrative Generation								
		Complete Sequence	60.66	56.67	60.71	58.62	57.14	50.55	60.53	55.09
		Statistical Summary	57.38	48.28	56.00	51.85	53.43	52.02	52.94	52.48
		Natural Language String	63.93	62.96	58.62	60.71	62.86	57.47	64.10	60.61
		Meta-Narrative (ours)	**70.49**	**65.22**	**60.00**	**62.50**	**64.86**	**61.11**	**67.48**	**64.14**

highest OOD performance. Improvements over best baselines: GLOBEM (+12.28% accuracy), LifeSnaps (+7.81% accuracy), MFAFY (+4.00% accuracy).

Narrative Generation Advantages: Prospective Narrative Generation consistently outperforms Binary Classification across all input representations. The largest improvement occurs on GLOBEM (69.40% vs 57.53% F1), demonstrating that generative formulations better leverage LLMs' contextual reasoning capabilities.

Published Benchmark Comparison: Our best GLOBEM configuration (67.40% OOD accuracy) substantially outperforms the published baseline (58.50% accuracy), representing a meaningful advancement in cross-distribution generalization for behavioral forecasting.

4.4 Analysis

Input Representation Impact Within the Prospective Narrative Generation formulation, Meta-Narrative consistently outperforms alternatives. Improvements over the next-best input representation: GLOBEM (+0.90% F1), LifeSnaps (+8.66% F1), MFAFY (+3.53% F1). The particularly strong improvement on LifeSnaps suggests contextual narratives are especially beneficial for physiological and psychological data requiring sophisticated temporal pattern interpretation.

Output Formulation Analysis The advantage of narrative generation is most pronounced with Meta-Narrative inputs. While other representations show 2-8% F1 improvements with narrative generation, Meta-Narrative shows 8-12% improvements, suggesting synergistic alignment with LLM capabilities.

Generalization Robustness To assess generalization stability, we analyze ID-OOD performance gaps. Meta-Narrative with Narrative Generation maintains small gaps in F1 scores across datasets (GLOBEM: 4.47%, LifeSnaps: 1.34%, MFAFY: -1.64%), while some baselines show large drops (e.g., Statistical Summary on LifeSnaps binary classification: 54.70% gap), indicating superior robustness against distribution shifts.

4.5 Bidirectional Validation

To rigorously validate the robustness of our approach, we perform comprehensive bidirectional evaluation, training models in both directions ($T \rightarrow T'$ and $T' \rightarrow T$) across all datasets. While the primary results for the **forward direction** ($T \rightarrow T'$)

Table 2: Average (μ) and standard deviation (σ) of OOD generalization performance across bidirectional experiments ($T \rightarrow T'$ and $T' \rightarrow T$) for Meta-Narrative input.

	GLOBEM		LifeSnaps		MFAFY	
Output Formulation	Acc ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)	Acc ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)	Acc ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
Binary Classification	55.10 ± 0.02	53.91 ± 3.62	57.87 ± 1.51	58.66 ± 1.99	64.53 ± 3.67	65.28 ± 0.32
Prospective Narrative Gen.	**68.08** $\pm$ 0.67	**67.92** $\pm$ 1.48	**69.31** $\pm$ 2.12	**68.94** $\pm$ 0.29	**67.67** $\pm$ 2.81	**64.07** $\pm$ 0.07

are detailed in Section 4.3, Table 2 offers a concise summary of performance statistics across *both* directions for Meta-Narrative input with both output formulations. The complete results for the **reverse direction** ($T' \rightarrow T$) are in Appendix A.11.

The bidirectional analysis reveals remarkable consistency patterns that strengthen our conclusions. **GLOBEM demonstrates exceptional stability**, with Binary Classification showing virtually identical performance across directions ($55.10 \pm 0.02\%$ accuracy), though F1 scores exhibit higher variance ($53.91 \pm 3.62\%$). For Prospective Narrative Generation, both accuracy and F1 remain highly consistent ($68.08 \pm 0.67\%$ and $67.92 \pm 1.48\%$, respectively), indicating robust bidirectional generalization.

LifeSnaps exhibits the strongest overall performance with Prospective Narrative Generation, achieving $69.31 \pm 2.12\%$ accuracy and remarkably stable F1 scores ($68.94 \pm 0.29\%$). The low F1 variance suggests excellent precision-recall balance across different temporal contexts. Interestingly, Binary Classification shows moderate directional sensitivity ($57.87 \pm 1.51\%$ accuracy), indicating that the choice of training direction matters more for discriminative than generative formulations.

MFAFY presents the most complex bidirectional behavior, with Binary Classification showing significant directional asymmetry ($64.53 \pm 3.67\%$ accuracy) but highly consistent F1 scores ($65.28 \pm 0.32\%$). This pattern reflects the temporal structure differences between one-semester (Year 2) and two-semester (Year 1) periods. Models trained on the more constrained Year 2 data achieve better generalization to Year 1 than vice versa, suggesting that training on focused, short-term data may lead to more transferable patterns. Despite this asymmetry, Prospective Narrative Generation maintains strong bidirectional performance ($67.67 \pm 2.81\%$ accuracy) with exceptional F1 consistency ($64.07 \pm 0.07\%$).

These bidirectional results provide compelling evidence that **ConText-LE's improvements stem from capturing fundamental data relationships rather than exploiting direction-specific biases**. The systematic advantages of narrative generation across all datasets and directions, combined with Meta-Narrative's consistent superiority, demonstrate robust generalization capabilities essential for real-world deployment where models must perform reliably across diverse temporal contexts.

4.6 LLM Architecture Ablation Study

To investigate how foundation model characteristics affect cross-distribution generalization, we evaluate three diverse LLMs on GLOBEM using our optimal configuration (Meta-Narrative + Prospective Narrative Generation): Llama 3.1 8B Instruct (our base model), Mistral-7B-Instruct-v0.3 (Mistral AI, 2024), and Falcon-7B (Almazrouei et al., 2023). The comparison includes two instruction-tuned models (Llama 3.1, Mistral-7B) and one base model (Falcon-7B), enabling assessment of both architectural differences and instruction tuning impact. All models undergo identical fine-tuning procedures as detailed in Appendix A.8.

Table 3: LLM architecture impact on GLOBEM cross-distribution generalization (Meta-Narrative + Prospective Narrative Generation).

	In-Distribution		Out-of-Distribution		ID-OOD Gap
LLM Architecture	Acc (%)	F1 (%)	Acc (%)	F1 (%)	F1 Gap (%)
Llama 3.1 8B Instruct	**73.99**	**73.87**	**67.40**	**69.40**	4.47
Mistral-7B-Instruct-v0.3	68.61	70.59	64.26	66.88	**3.71**
Falcon-7B	62.78	64.68	56.15	59.66	5.02

Table 3 presents comprehensive performance metrics across ID and OOD settings.

- **Instruction Tuning Criticality**: Instruction-tuned models substantially outperform the base model (Llama vs. Falcon: +9.74% F1; Mistral vs. Falcon: +7.22% F1), demonstrating that instruction tuning is essential for interpreting contextual behavioral narratives effectively.
- **Context Length Advantages**: Llama 3.1's extended context (128K tokens) compared to Mistral-7B (32K) and Falcon-7B (4K) enables superior understanding of long-term temporal patterns within Meta-Narratives, contributing to its performance advantage.
- **Generalization Stability**: Mistral-7B exhibits the smallest ID-OOD gap (3.71%), followed by

Llama 3.1 (4.47%), while Falcon-7B shows the largest gap (5.02%). This indicates that architectural efficiency and instruction tuning contribute more to stable generalization than raw parameter count.

These results confirm that while ConText-LE provides an effective framework for behavioral forecasting, the choice of foundation model significantly impacts cross-distribution performance. Instruction tuning, extended context length, and diverse pre-training data emerge as key architectural factors for robust behavioral pattern interpretation.

4.7 Key Findings

Our comprehensive evaluation establishes several critical findings:

I. **Meta-Narrative Superiority**: Consistently outperforms alternative text representations across all datasets and output formulations, with F1 improvements ranging from 0.90% (GLOBEM) to 8.66% (LifeSnaps) over the next-best input representation.

II. **Generative Formulation Advantages**: Prospective Narrative Generation systematically outperforms Binary Classification across all configurations. The benefits are most pronounced with Meta-Narrative inputs, showing 11.87% (GLOBEM) to 11.99% (LifeSnaps) F1 improvements.

III. **Cross-Distribution Robustness**: Meta-Narrative with Narrative Generation maintains small ID-OOD gaps (1.34% to 4.47% F1) and demonstrates consistent bidirectional performance, validating that improvements capture fundamental behavioral relationships rather than temporal artifacts.

IV. **Foundation Model Dependencies**: LLM architecture choice significantly impacts generalization performance. Instruction tuning provides substantial benefits (+7.22% to +9.74% F1), while extended context length and diverse pre-training enhance temporal pattern interpretation.

V. **Benchmark Advancement**: Achieves substantial improvements over published baselines (e.g., +14.90% accuracy over GLOBEM's published OOD baseline), demonstrating practical viability for reliable cross-distribution behavioral forecasting.

These findings establish ConText-LE as a significant advancement in generalizable LE data modeling, providing both theoretical insights into LLM-based contextual representation learning and practical improvements for behavioral prediction systems deployed across diverse temporal and demographic contexts.

5 Discussion

Our comprehensive evaluation demonstrates that contextual narrative representations fundamentally improve cross-distribution generalization in longitudinal experiential data modeling. Three key insights emerge from this work.

First, **contextual narrative representations are crucial for generalization**. The consistent superiority of Meta-Narrative over simpler encodings across all datasets and metrics indicates that semantically rich representations capturing complex feature relationships are essential for robust cross-distribution performance. This aligns with recent NLP advances showing that contextually rich inputs significantly improve complex reasoning tasks (Wei et al., 2023; Wang et al., 2022a).

Second, **generative formulations enhance cross-domain transfer**. Prospective Narrative Generation's systematic advantages over Binary Classification suggest that allowing models to generate nuanced predictions fosters better reasoning about complex behavioral patterns. This generative capability facilitates adaptation to novel contexts, echoing broader NLP findings where generative approaches often excel in complex reasoning scenarios (Kojima et al., 2023).

Third, **representational alignment with LLM capabilities is critical**. The synergistic effects observed when combining Meta-Narrative inputs with Narrative Generation outputs indicate that optimizing both input representation and output task to match LLMs' strengths in contextual understanding unlocks robust generalization. This holistic alignment, consistent across diverse datasets and bidirectional evaluations, confirms that ConText-LE captures fundamental behavioral relationships rather than exploiting dataset-specific artifacts.

These findings establish ConText-LE as a principled framework for leveraging LLMs' contextual understanding in behavioral forecasting, with clear implications for developing more reliable AI systems in sensitive domains like mental health and education.

6 Limitations and Future Work

While ConText-LE demonstrates significant advances in cross-distribution generalization for longitudinal experiential data, several important limitations point to valuable directions for future research.

6.1 Current Limitations

External LLM Dependency A critical limitation is the reliance on GPT-4o for Meta-Narrative generation, target creation, and prediction extraction. This dependency creates deployment challenges: (1) external API costs and latency constraints, (2) potential quality variations across LLM versions, (3) limited control over representation consistency, and (4) barriers for privacy-sensitive or resource-constrained environments.

Failure Mode Analysis Qualitative analysis reveals systematic failure patterns: (1) over-reliance on recent temporal patterns without broader contextual integration, (2) difficulty resolving conflicting behavioral signals (e.g., high stress but stable mood), (3) limited domain-specific knowledge affecting interpretation of context-dependent events (e.g., academic examination periods, clinical interventions).

Computational Requirements Despite using LoRA for efficient fine-tuning, the approach requires substantial computational resources for both training and inference. The multi-stage processing pipeline introduces latency that may limit real-time deployment scenarios, while GPU requirements may restrict accessibility for practitioners with limited resources.

Limited Mechanistic Understanding The "black-box" nature of LLMs limits insight into causal mechanisms behind improved generalization. This constrains systematic improvement based on principled understanding rather than empirical exploration, and prevents clear identification of which narrative components most critically contribute to performance.

Domain and Scale Limitations Evaluation focuses on mental health and education domains with moderate-scale datasets. Generalizability to other LE data contexts (e.g., physical health, workplace performance), larger datasets, or more severe distribution shifts (e.g., cross-cultural generalization) remains unverified.

6.2 Future Research Directions

Reducing External Dependencies Priority should be given to developing self-contained approaches that eliminate GPT-4o dependency. Promising directions include: (1) training specialized distilled models for representation generation (Hinton et al., 2015), (2) end-to-end architectures incorporating representation learning directly into forecasting models through multi-task objectives (Collobert and Weston, 2008), (3) domain-specific pre-training strategies for LE data (Gururangan et al., 2020).

Interpretability and Mechanistic Understanding Future work should incorporate systematic interpretability analyses to understand generalization mechanisms: (1) ablation studies varying narrative components systematically, (2) attention flow analyses tracking information propagation (Abnar and Zuidema, 2020), (3) probing studies identifying linguistic features correlating with performance (Hewitt and Manning, 2019), (4) development of more transparent models maintaining contextual benefits while offering interpretability.

Computational Efficiency Research should explore efficiency optimizations specifically for LE data: (1) knowledge distillation for model compression (Hinton et al., 2015), (2) adaptive architectures combining lightweight and powerful components, (3) quantization and pruning techniques (Dettmers et al., 2022; Frankle and Carbin, 2019), (4) specialized hardware-software co-design for behavioral forecasting workloads.

Broader Evaluation and Robustness Extending evaluation scope is crucial: (1) diverse LE data domains and larger datasets, (2) cross-cultural and cross-demographic generalization studies, (3) more severe distribution shifts and longer temporal gaps, (4) comprehensive comparisons with multimodal approaches and specialized time series architectures.

Ethical and Privacy Considerations Future development must integrate ethical considerations: (1) privacy-preserving narrative representations minimizing identifiable information, (2) fairness analysis across demographic groups, (3) bias mitigation in cross-population generalization, (4) clear guidelines for appropriate use cases and consent frameworks, (5) interdisciplinary collaboration with domain experts and ethicists.

Narrative Quality and Consistency Systematic approaches to narrative optimization should be developed: (1) specialized metrics for narrative quality in behavioral contexts, (2) consistency checking mechanisms detecting spurious correlations, (3) fact verification techniques adapted for behavioral narratives (Thorne et al., 2018), (4) coherence modeling for temporal behavioral descriptions (Iter et al., 2020).

Despite these limitations, ConText-LE represents a significant step toward more generalizable LE data modeling by demonstrating the effectiveness of contextual narrative representations. The identified limitations offer concrete directions for advancing the field toward more reliable, efficient, and ethically sound behavioral forecasting systems.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *Preprint*, arXiv:2005.00928.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

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.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

S. Bae, D. Ferreira, B. Suffoletto, J. C. Puyana, R. Kurtz, T. Chung, and A. K. Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):5.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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, and 95 others. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 1293–1304, New York, NY, USA. Association for Computing Machinery.

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. *Preprint*, arXiv:2310.04948.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *Preprint*, arXiv:2308.08469.

P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, M. Goel, and A. K. Dey. 2021. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Trans. Comput.-Hum. Interact.*, 28(1):3:1–3:41.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *Preprint*, arXiv:2110.02861.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)*, pages 1–8.

Charles J. Fillmore. 2006. Chapter 10 frame semantics. In Dirk Geeraerts, editor, *Cognitive Linguistics: Basic Readings*, pages 373–400. De Gruyter Mouton, Berlin, New York.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *Preprint*, arXiv:1803.03635.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. *Preprint*, arXiv:2310.07820.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online. Association for Computational Linguistics.

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. 2023. Time-llm: Time series forecasting by reprogramming large language models. *Preprint*, arXiv:2310.01728.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *Preprint*, arXiv:2401.06866.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Santosh Kumar, Gregory D. Abowd, William T. Abraham, Mustafa al'Absi, J. Gayle Beck, Duen Horng Chau, Tyson Condie, David E. Conroy, Emre Ertin, Deborah Estrin, Deepak Ganesan, Chia-Fang Lam, Benjamin Marlin, Charles B. Marsh, Susan A. Murphy, Inbal Nahum-Shani, Kevin Patrick, James M. Rehg, Moinul Sharmin, and 5 others. 2015. Center of excellence for mobile sensor data-to-knowledge (md2k). *Journal of the American Medical Informatics Association (JAMIA)*, 22(6):1137–1142.

Reed Larson and Mihaly Csikszentmihalyi. 1983. The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*, 15:41–56.

Xianling Li, Xiaoyan Zhu, Xiaofei Zhu, Yunkai Ji, and Xiangnan Tang. 2020. Student academic performance prediction using deep multi-source behavior sequential network. In *Advances in Knowledge Discovery and Data Mining. PAKDD 2020*, volume 12084 of *Lecture Notes in Computer Science*, pages 570–582. Springer, Cham.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *Preprint*, arXiv:2305.15525.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Preprint*, arXiv:1310.4546.

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

David C. Mohr, Francisca Azocar, Andrea Bertagnolli, Tanzeem Choudhury, Paul Chrisp, Richard Frank, Henry Harbin, Trina Histon, Debra Kaysen, Camille Nebeker, Derek Richards, Stephen M. Schueller, Nickolai Titov, John Torous, Patricia A. Areán, and Banbury Forum on Digital Mental Health. 2021. Banbury forum consensus statement on the path forward for digital mental health treatment. *Psychiatric Services*, 72(6):677–683. Epub 2021 Jan 20.

David C. Mohr, Michelle Zhang, and Stephen M. Schueller. 2017. Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13:23–47.

Samuel T. Moulton and Stephen M. Kosslyn. 2009. Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1273–1280.

Ehsan Nemati, Xiangyu Xu, Varun Nathan, Kaveh Vatanparvar, Tanzima Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jing Kuang, and Anhong Gao. 2022. Ubilung: Multi-modal passive-based lung health assessment. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 551–555. IEEE.

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

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

M. Rabbi, P. Klasnja, T. Choudhury, A. Tewari, and S. Murphy. 2019. Optimizing mhealth interventions with a bandit. In H. Baumeister and C. Montag, editors, *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, Studies in Neuroscience, Psychology and Behavioral Economics, pages 277–291. Springer International Publishing, Cham.

Sohrab Saeb, Michelle Zhang, Christopher J. Karr, Stephen M. Schueller, Molly E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7):e175.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Daniel L. Schacter, Donna Rose Addis, and Randy L. Buckner. 2008. Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences*, 1124:39–60.

Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Arthur A. Stone and Saul Shiffman. 1994. Ecological momentary assessment (ema) in behavioral medicine. *Annals of Behavioral Medicine*, 16(3):199–202.

Chenxi Sun, Yaliang Li, Hongyan Li, and Shenda Hong. 2023. Test: Text prototype aligned embedding to activate llm's ability for time series. *Preprint*, arXiv:2308.08241.

Technology Innovation Institute. 2023. Falcon-7b model card. *Hugging Face*. Retrieved from https://huggingface.co/tiiuae/falcon-7b.

Nguyen T. Thach, Patrick Habecker, Anika R. Eisenbraun, W. Alex Mason, Kimberly A. Tyler, Bilal Khan, and Hau Chan. 2025. Muhboost: Multi-label boosting for practical longitudinal human behavior modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Accepted. Available at https://openreview.net/pdf?id=BAelAyADqn.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile sensing and support for people with depression: A pilot trial in the wild. *JMIR mHealth and uHealth*, 4(3):e111.

Jianing Wang, Wenkang Huang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Xiang Li, and Ming Gao. 2022a. Knowledge prompting in pre-trained language model for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3177, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Rui Wang, Wenbo Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):43:1–43:26.

Weichen Wang, Suranga Nepal, John F. Huckins, Leidy Hernandez, Vlado Vojdanovski, Dianne Mack, Jamie Plomp, Akshay Pillai, Mariko Obuchi, Ashley Dasilva, Eamon Murphy, Emma Hedlund, Christopher Rogers, Morgan Meyer, and Andrew Campbell. 2022b. First-gen lens: Assessing mental health of first-generation students across their first year at college using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):95. Epub 2022 Jul 7.

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 preprint*. ArXiv:2201.11903 [cs].

X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, J. Mankoff, and A. K. Dey. 2019. Leveraging routine behavior and contextually-filtered features for depression detection among college students. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, volume 3, pages 116:1–116:33.

Xiangyu Xu, Prathyusha Chikersal, Jessica M. Dutcher, Yeganeh S. Sefidgar, Woojin Seo, Michael J. Tumminia, Daniel K. Villalba, Susan Cohen, Kelsy G. Creswell, John D. Creswell, Ali Doryab, Paula S. Nurius, Elizabeth Riskin, Anind K. Dey, and Jennifer Mankoff. 2021a. Leveraging collaborative-filtering for personalized behavior modeling: A case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):41:1–41:27.

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. 2023a. 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, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021b. Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1).

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 E. Morris, Eve Riskin, Jennifer Mankoff, and Anind K. Dey. 2023b. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. *Preprint*, arXiv:2211.02733.

Hao Xue and Flora D. Salim. 2023. PromptCast: A New Prompt-based Learning Paradigm for Time Series

Forecasting. *arXiv preprint*. ArXiv:2210.08964 [cs, math, stat].

Sofia Yfantidou, Christina Karagianni, Stelios Efstathiou, and 1 others. 2022. Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data*, 9(1):663.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. 2025. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *Preprint*, arXiv:2502.04395.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20.

Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One fits all:power general time series analysis by pretrained lm. *Preprint*, arXiv:2302.11939.

A Appendix

A.1 Detailed Problem Formulation

This section provides a more detailed and formal specification of the problem formulation for generalizable LE data forecasting within the ConText-LE framework, expanding upon Section 3.

We consider LE data collected from a set of N individuals over a total observation period T, spanning K weeks. Data is recorded at a daily granularity, resulting in T_{total} daily time steps, where $T_{total} = K \times 7$.

For each individual $i \in \{1, \dots, N\}$ and each daily time step $j \in \{1, \dots, T_{total}\}$, we have a feature vector $x_{i,j} \in \mathbb{R}^D$, where D is the total number of features. These features $x_{i,j}$ encompass diverse modalities and types (e.g., numerical sensor readings, categorical logs, free-text self-reports).

The forecasting task is framed using a sliding window approach with a window size of k weeks. For each individual i, we extract overlapping input sequences. An input sequence starting at week s (where $s \in \{1, \dots, K-k\}$) corresponds to the raw data $\{x_{i,j}\}$ for all daily time steps j within the period spanning week s through week $s+k-1$. Let $J_{s,s+k-1}$ denote the set of daily time step indices corresponding to weeks s through $s+k-1$. The raw data for an input sequence is thus $\{x_{i,j} \mid j \in J_{s,s+k-1}\}$.

This raw data sequence is transformed into a textual representation, denoted as $X^{\text{text-rep}}_{i,s\dots s+k-1}$. This transformation is performed using one of the four strategies detailed in Section **??**: Complete Sequence, Statistical Summary Encoding, Natural Language String Encoding, or Meta-Narrative. The specific format of $X^{\text{text-rep}}_{i,s\dots s+k-1}$ depends on the chosen strategy.

The target for the forecasting task is defined for the week immediately following the input window, i.e., week $s+k$. We investigate two output formulations:

1. **Binary Label Target ($y^{\text{binary}}_{i,s+k}$):** A binary value indicating a specific state (e.g., depression: high/low; engagement: yes/no) for individual i at week $s+k$, i.e., $y^{\text{binary}}_{i,s+k} \in \{0, 1\}$.

2. **Prospective Narrative Target ($y^{\text{text}}_{i,s+k}$):** A natural language sequence describing or aligned with the actual state of individual i at week $s+k$; used as the target for text generation.

The problem is to train an LLM to learn a mapping function f from the textual input representation $X^{\text{text-rep}}_{i,s\ldots s+k-1}$ to either the binary label target $y^{\text{binary}}_{i,s+k}$ (for the Binary Classification formulation) or the prospective narrative target $y^{\text{text}}_{i,s+k}$ (for the Prospective Narrative Generation formulation):

$$f : X^{\text{text-rep}}_{i,s\ldots s+k-1} \rightarrow \begin{cases} y^{\text{binary}}_{i,s+k} & \text{(binary classification)} \\ y^{\text{text}}_{i,s+k} & \text{(prospective narrative)} \end{cases}$$

The primary objective is to learn an f that exhibits strong generalization performance when applied to data from a distinct period or cohort (T') not seen during training on data from source period T. Evaluation metrics (Accuracy, Precision, Recall, F1) are computed based on the binary forecast extracted from the model's output (either directly from the classification head or inferred from the generated narrative).

A.2 Examples of Textual Representations

This section provides illustrative examples of the four textual representation strategies discussed in Section 3. For demonstration purposes, we use a simplified hypothetical k-week input sequence involving a few representative features (e.g., Steps, Sleep Duration, Mood). Note that actual generated texts using GPT-4o may vary in phrasing but adhere to the defined format and content goals for each strategy.

Hypothetical k-week Raw Data Excerpt (Imagine raw data for 2 weeks, with daily values for Steps, Sleep, and Mood)

- **Complete Sequence Example:** *Week 1 started with the user taking 500 steps on Day 1, followed by 1200 steps on Day 2. Sleep was 7 hours on Day 1 and 8.5 hours on Day 2. Mood was reported as 3 on both days. Day 3 data is missing for all features. Day 4 had 800 steps, 7.8 hours of sleep, and mood was 4... The second week began with 1500 steps on Day 8, sleep was 7.2 hours, and mood was 3, continuing through Day 14...*

- **Statistical Summary Encoding Example:** *Statistical summary over the k-week period: Steps: "avg": 1050, "std": 350, "min": 500, "max": 1500 steps. Sleep Duration: "avg": 7.5, "std": 0.6, "min": 6.0, "max": 8.5 hours. Mood: "avg": 3.5, "std": 0.5, "min": 3, "max": 4 out of 5.*

- **Natural Language String Encoding Example:** *Steps: ["500", "1200", "300", "800", ..., "1500", ...]. Sleep Duration: ["7.0", "8.5", "400", "7.8", ..., "7.2", ...]. Mood: ["3", "3", "500", "4", ..., "3", ...]. (Note: Specific formatting like brackets, and commas, representation may vary slightly based on prompt design, but the core structure of listing values chronologically per feature is consistent.)*

- **Meta-Narrative Example:** *Over the past k weeks, the user's activity levels showed moderate fluctuation with an overall increasing trend towards the end of the period. Sleep patterns remained relatively stable, averaging around 7.5 hours per night, though some variability was noted. Mood reports were generally consistent, hovering between 3 and 4, without significant sharp declines or improvements.*

These examples illustrate the different ways each strategy encodes the same underlying LE data into a textual format for processing by the LLM. The Complete Sequence offers maximal detail, Statistical Summary provides aggregates, Natural Language String gives a structured temporal listing, and the Meta-Narrative provides a high-level interpretation.

A.3 Output Formulations for Forecasting

ConText-LE investigates two distinct ways to formulate the prediction target and task for the LLM, hypothesizing that a generative narrative output aligns better with LLMs' core capabilities for generalizable LE data modeling than traditional classification.

Binary Classification Formulation In this traditional formulation, the prediction target is a single binary label $y^{\text{binary}}_{i,s+k}$ (e.g., 0 or 1, representing "low depression" or "high depression"). We adapt a pre-trained LLM by replacing its original language modeling head with a Sequence Classification head. The model is fine-tuned in a supervised manner, mapping the textual input representation ($X^{\text{text-rep}}_{i,s\ldots s+k-1}$) directly to the binary target label ($y^{\text{binary}}_{i,s+k}$). The loss function is cross-entropy, calculated between the predicted binary label distribution and the one-hot encoded true label. During inference, the fine-tuned LLM outputs a probability distribution over the two classes, and the class with the highest probability is taken as the final forecast.

Prospective Narrative Generation Formulation In this formulation, inspired by cognitive processes of integrated forward-looking assessment (Moulton and Kosslyn, 2009; Schacter et al., 2008), the forecasting task is reframed as a language generation problem. The prediction target is a natural language text sequence, the **prospective narrative** $y^{\text{text}}_{i,s+k}$, which implicitly encodes the predicted future state for week $s+k$. The pre-trained LLM is fine-tuned using a causal language modeling objective to generate this target narrative based on the textual input representation ($X^{\text{text-rep}}_{i,s\ldots s+k-1}$).

This approach builds on recent findings in NLP that generative formulations can be more effective than discriminative ones for complex reasoning tasks (Wei et al., 2023; Kojima et al., 2023; Wang et al., 2022a). By allowing the model to generate a narrative prediction rather than forcing a binary decision, we enable it to articulate subtle contextual relationships and degrees of certainty that might be lost in classification. For LE data in particular, where interpretation depends heavily on contextual factors beyond statistical patterns, this generative approach may better leverage LLMs' pre-trained understanding of how features interact in complex human behaviors.

To obtain these training targets ($y^{\text{text}}_{i,s+k}$), we leverage GPT-4o. For each k-week input sequence from the training data, paired with its ground truth actual state or outcome for the subsequent week ($y^{\text{actual}}_{i,s+k}$), GPT-4o is prompted to generate a narrative reflection on the past k-week trajectory that aligns with or anticipates the known actual state for week $s+k$. This process is detailed in Appendix A.5. During inference, the fine-tuned LLM generates a prospective narrative based on the input.

A.4 Input Textualization Prompts

LLM Prompt for Summary This prompt guides the model to generate a concise, human-like behavioral interpretation that highlights key psychological trends—such as shifts in motivation, confidence, and future orientation—across a 4-week period. Rather than quoting student responses, it encourages abstraction and synthesis, allowing the model to infer meaningful behavioral patterns.

> **System Prompt – Statistical Summary**
>
> You are an expert in behavioral analysis. Your task is to generate a concise, natural-sounding 3–4 line summary of a student's 4-week behavioral log. The log reflects the student's motivation, attitude, confidence, and future orientation. Identify high-level trends and patterns in their reflections without quoting directly. Focus on behaviorally meaningful changes or consistencies.

LLM Prompt for Complete Sequence This prompt presents the model with a detailed, temporally structured sequence of student reflections organized by week and day. It preserves the full chronology of responses, allowing the model to track behavioral progression over time and identify week-to-week shifts in motivation, engagement, or outlook based on the specific timing and context of student inputs.

> **System Prompt – Complete Sequence**
>
> You are an expert in prompt engineering and behavioral analysis. You are given a student's 4-week chronological reflection log, structured by week and day (e.g., "Week 1:", "Monday:"), with entries for pre-lecture anticipation, post-lecture reflection, confidence, and future orientation. Your task is to write a clear and effective system prompt that can be used to instruct a language model to analyze this type of structured input and identify behavioral trends over time.

System Prompt Design for Natural Language String This prompt was developed to reflect the flattened, theme-based organization of the input, where responses are grouped by behavioral dimensions such as confidence or motivation rather than by time. The instruction explicitly mentions that each segment is prefixed by a label indicating its thematic category. The prompt guides the model to interpret patterns across these categories without being constrained by temporal order, and to infer meaningful behavioral shifts or consistencies across the entire 4-week period based on thematic clustering rather than day-to-day variation.

System Prompt – Natural Language String

You are an expert in prompt engineering and behavioral interpretation. You are provided with a theme-based summary of student reflections over four weeks. Each segment is labeled by behavioral category (e.g., confidence, motivation, peer comparison). Your task is to generate a system prompt that can instruct a language model to interpret this type of grouped input and produce a behavioral analysis based on observed trends across these categories.

LLM Prompt for Textual Meta-Narrative Generation For the Meta-Narrative approach specifically, we implement a two-stage prompting process inspired by recent advances in multi-step reasoning techniques (Wei et al., 2023; Kojima et al., 2023):

1. **Feature Pattern Analysis**: First, GPT-4o analyzes each feature's temporal trajectory separately, identifying significant patterns, trends, and anomalies. The prompt includes domain-specific context (e.g., university student behaviors, mental health indicators) to guide interpretation. This step leverages the LLM's ability to detect statistical patterns within individual features, similar to how contextualized language models learn to represent individual tokens within their local context (Peters et al., 2018).

2. **Contextual Narrative Synthesis**: Second, GPT-4o integrates these individual feature analyses into a coherent narrative that emphasizes inter-feature relationships and contextual interpretations grounded in human behavior patterns. This step parallels how contextualized language models integrate token-level representations into coherent sentence-level semantics (Devlin et al., 2019; Liu et al., 2024).

This two-stage process transforms multi-dimensional time-series data into contextually rich narratives, effectively capturing cross-feature dependencies and temporal dynamics that might be lost in simpler representations. The Meta-Narrative approach is designed to leverage LLMs' pre-trained understanding of how events and behaviors relate to each other in meaningful ways, creating inputs that are semantically coherent and contextually grounded. The LLM prompt is give below.

System Prompt – Meta-Narrative

You are an expert behavioral analyst tasked with evaluating a student's weekly behavioral reflections over a 4-week course. The data includes daily pre- and post-lecture thoughts, confidence levels, peer comparisons, and future-oriented reflections.

Your objective is to analyze the evolution of the student's behavior and mindset across the 4 weeks. In your response:

- Identify and describe specific behavioral trends, such as shifts in confidence, motivation, or engagement.
- Reference specific weeks (e.g., "In Week 1...", "By Week 3...").
- Use precise language to describe changes, such as "X increased by Week 2", "Y decreased from Week 1 to Week 4", or "Z remained consistent until Week 3".
- Avoid vague terms like "overall" or "in general" to ensure analytical precision.
- Provide a concise, natural, and evidence-based analysis in 3–4 sentences.
- Exclude any personal or identifying information from the response.

A.5 LLM Prompt for Prospective Narrative Generation

System Prompt – Prospective Narrative Generation

You are an expert behavioral analyst. A student's weekly behavioral reflections over a 4-week course are provided below, including daily pre- and post-lecture thoughts, confidence levels, peer comparisons, and future-oriented reflections:

`{input_text}`

The student's behavior is labeled as `'{output_label}'`.

Write a clear, natural-language expert explanation — just a single 3–4 sentence paragraph explaining the behavioral trends that support the label. Be concise and insightful, as if communicating with another expert. Avoid vague terms like "overall" or "in general," and exclude any personal or identifying information.

A.6 LLM Prompt for Prediction Extraction from Generated Narrative

System Prompt – Prediction Extraction

You are a student engagement expert. Based on the behavioral reasoning below, classify the student's confidence level as either `High` or `Low`. You must choose one. No explanation.
Reasoning:
`{reasoning_text}`
Output only: `High` or `Low`.

A.7 Design Principles for Contextual Understanding

The ConText-LE framework's design is guided by three core principles from NLP research on contextual representation learning:

- **Semantic Coherence**: The Meta-Narrative representation transforms discrete time-series data into a coherent narrative with integrated semantic meaning. This approach draws on findings that LLMs perform better when information is presented in coherent, semantically rich formats (Wang et al., 2022a; Shwartz et al., 2020). By constructing a narrative that emphasizes relationships between features, we better leverage LLMs' pre-trained understanding of how elements gain meaning through their context.
- **Generative Expression**: The Prospective Narrative Generation formulation aligns with recent work showing that generative approaches often outperform discriminative ones for complex reasoning tasks (Wei et al., 2023; Kojima et al., 2023). By generating narratives rather than binary labels, the model can express nuanced predictions with implicit uncertainty and conditional reasoning that better captures the complexity of human behavioral forecasting.
- **Hierarchical Processing**: The two-stage process for Meta-Narrative generation applies the hierarchical processing principles from successful NLP architectures. Similar to how models like BERT (Devlin et al., 2019) build higher-level representations from lower-level ones, our approach first analyzes individual features before synthesizing them into an integrated narrative, enabling better capture of both local patterns and global relationships.

These design principles are motivated by the observation that LLMs excel at tasks when the representation and processing align with how they were pre-trained to understand language. By structuring both input representations and output formulations to leverage LLMs' core capabilities in contextual understanding and narrative generation, we hypothesize improved cross-distribution robustness compared to approaches that treat LE data as simple statistical patterns.

A.8 Implementation Details

External LLM Usage (GPT-4o) ConText-LE leverages the advanced capabilities of GPT-4o (OpenAI, 2024) for several crucial steps in the pipeline, particularly during data preparation for training and output processing for evaluation. These steps are performed via API calls using carefully designed prompts.

- **Textual Representation Generation**: GPT-4o transforms raw k-week LE data sequences into two textual representation strategies—**Statistical Summary** and **Meta-Narrative**—as described earlier. For the **Meta-Narrative** specifically, this involves a two-stage process: *Feature Pattern Analysis* followed by *Contextual Narrative Synthesis*, implemented through sequential prompting with context carried forward between steps.
- **Target Prospective Narrative Generation**: For the Prospective Narrative Generation formulation, GPT-4o generates the target narrative texts ($y^{\text{text}}_{i,s+k}$) during training data preparation. The prompt includes the input sequence and ground truth outcome, instructing GPT-4o to generate a narrative that contextually aligns with that outcome.
- **Forecast Extraction from Narratives**: For evaluation of the Prospective Narrative Generation formulation, GPT-4o extracts binary forecasts from generated narratives. This enables quantitative comparison with ground truth labels and other methods. To ensure consistency, we use structured zero-shot prompting with explicit instructions to identify the implied prediction within the generated narrative.

The reliance on this external LLM for these processing steps represents a practical consideration in our current implementation and is discussed as a limitation in Section 6.

Fine-tuning Process We employ parameter-efficient fine-tuning (PEFT) using LoRA (Hu et al.,

2021) to adapt the LLM while keeping most of its parameters frozen. This approach reduces computational requirements while allowing the model to adapt to the specialized LE data domain. The fine-tuning process differs based on the output formulation in Binary Classification and Prospective Narrative Generation. Detailed fine-tuning hyperparameters for both formulations are provided in Appendix A.8.

Inference Process During inference on unseen k-week data sequences, the same input transformation pipeline is applied using the chosen textual representation strategy. The fine-tuned LLM then processes this textual input.

- **Binary Classification**: The LLM with the classification head directly outputs the predicted binary label (0 or 1).
- **Prospective Narrative Generation**: The LLM generates a sequence of tokens constituting the predictive prospective narrative. For this formulation, we use a temperature of 0.7 and top-p sampling with p=1.0 to balance deterministic prediction with narrative richness. We set a maximum generation length of 300 tokens and apply a frequency penalty of 0.5 to avoid redundant text.

For quantitative evaluation, the predictive narrative output from the Prospective Narrative Generation formulation requires an additional step to obtain a binary forecast comparable to ground truth. We use GPT-4o to extract a textual binary label from the predictive narrative, using a carefully designed prompt that focuses on identifying the implied forecast within the generated text. The prompt used for this extractive task is given in Appendix A.6.

LLM Fine-tuning Configuration For all experiments, we utilize Llama 3.1 8B Instruct (Grattafiori et al., 2024) as the base LLM, selected for its strong performance on language understanding and generation tasks while remaining computationally efficient. We employ parameter-efficient fine-tuning (PEFT) using LoRA (Hu et al., 2021) to adapt the LLM while keeping most of its parameters frozen. This approach reduces computational requirements while allowing the model to adapt to the specialized LE data domain. The fine-tuning process differs based on the output formulation:

- **Binary Classification**: The LLM is fine-tuned with a Sequence Classification head added on top of its last hidden state. LoRA is applied to the query, key, and value projection matrices in each transformer layer, with a rank of 8. The model learns to map the input sequence to the binary label.
  - Parameter-efficient fine-tuning: LoRA (Hu et al., 2021) with:
    - Rank: 32
    - Alpha: 16
    - Target modules: All attention modules in the language model
  - Training objective: Causal language modeling with teacher forcing
  - Optimizer: paged-AdamW-8bit
  - Learning rate: 1e-5 with cosine decay schedule
  - Warmup-ration: 0.1
  - Batch size: 8
  - Training epochs: 20
  - Mixed precision: bfloat16
- **Prospective Narrative Generation**: The LLM is fine-tuned using a causal language modeling objective. LoRA is applied to the same projection matrices but with a rank of 16 to accommodate the more complex generation task. The model learns to generate the output narrative token by token.

- Parameter-efficient fine-tuning: LoRA (Hu et al., 2021) with:
  - Rank: 32
  - Alpha: 16
  - Target modules: All attention modules in the language model
- Training objective: Causal language modeling with teacher forcing
- Optimizer: paged-AdamW-8bit
- Learning rate: 1e-5 with cosine decay schedule
- Warmup-ration: 0.1
- Batch size: 8
- Training epochs: 20
- Mixed precision: bfloat16

Training Hardware Training was conducted on 8 × NVIDIA A40 GPUs (48GB each) with distributed data parallelism.

A.9 Datasets

We utilize the following LE datasets, selected for their relevance to health and behavioral forecasting and their suitability for evaluating challenging generalization across different cohorts and time periods:

- **GLOBEM (Xu et al., 2023a):** This is a widely used benchmark for longitudinal human behavior modeling and generalization. It comprises data collected from 497 unique participants across two institutions over four years (Year 1 & 2 from Institution A, Year 3 & 4 from Institution B), resulting in 661 person-years of data after initial preprocessing steps. Institutions A (pre-COVID) and B (post-COVID) represent distinct cohorts and time periods, with surveys including PHQ-4, BDI-II, and PANAS for depression assessment. We utilize a subset of 15 features based on prior work (Xu et al., 2023a; Thach et al., 2025; Kim et al., 2024), derived from mobile sensing data sources, including Location (variance, entropy, travel distance, duration of stay), Phone Usage (unlock counts, stats), Bluetooth (scan counts, unique devices), Call (duration stats, missed call count), Physical Activity (steps, active/sedentary duration), and Sleep (duration, episode stats). For the main evaluation, we use data from Years 1 & 2 from Institution A (344 person-years) for training and data from Years 3 & 4 from Institution B (317 person-years) for cross-cohort and cross-temporal generalization testing. Each person-year of data represents a 10-week observation period from which 6 sequences are generated using a 4-week sliding window predicting the subsequent week. This results in a training set of approximately 2226 sequences and a test set of approximately 2023 sequences. The task is binary mental health prediction based on a threshold applied to survey scores, resulting in a nearly balanced distribution.

- **LifeSnaps (Yfantidou et al., 2022):** This is a multi-dimensional LE dataset initially collected from 71 participants over 4 months, capturing unobtrusive snapshots of real-world human behavior in the wild. Data sources include Fitbit sensing data (e.g., activity, sleep, stress, heart rate), EMAs (e.g., mood, context), and validated surveys (e.g., psychological traits). The dataset includes over 35 distinct data types. For this work, we use a subset of relevant features from these modalities to predict a binary anxiety level in the week subsequent to a k=1 week observation window. After initial preprocessing steps, including filtering participants with significant missing values, a subset of participants was used for the evaluation splits. The specific cross-distribution split for evaluation involves training on data from 26 participants collected during the first 2 months of the study period and testing on data from 13 disjoint participants collected during the last 2 months, assessing cross-temporal and cross-participant generalization within the study cohort. Using a k=1 week window over these approximately 8-week periods yields a training set of approximately 112 sequences and a test set of approximately 64 sequences. This dataset serves to further validate cross-study generalization within the mental health domain using a different dataset structure, population, and data collection protocol.

- **MFAFY (Hayat et al., 2024a,b; Thach et al., 2025):** The Messages From A Future You (MFAFY) dataset captures aspects of first-year college students' academic journey over three consecutive semesters spanning two academic years (Year 1: Semesters 1 & 2; Year 2: Semester 3). It is a high-dimensional dataset comprising non-cognitive (28 dimensions, qualitative, e.g., motivation, engagement), cognitive (41 dimensions, quantitative, e.g., assessment scores), and background factors (9 dimensions, static qualitative, e.g., academic meta-information). For forecasting student behavioral engagement, we predict a student's lecture-related engagement status (binary: high-/low) in the subsequent week, using a k=4 week observation window. Input features use only relevant non-cognitive dimensions. The binary target is derived by comparing the average of relevant non-cognitive features during weeks s through $s+k-1$ with their average during week $s+k$. This task results in a nearly balanced binary distribution. For evaluation, the cross-year generalization split consists of a training set using data from 61 subjects in Year 1 (Semesters 1 & 2) and a test set using data from 35 subjects in Year 2 (Semester 3). Each subject-year/semester of data represents a 15-week observation period from which 10 sequences are generated using a 4-week sliding window predicting the subsequent week. This results in a training set of approximately 610 sequences and a test set of approximately 350 sequences.

For all datasets, train/test splits are carefully created to ensure strict separation of data from different cohorts or time periods for generalization evaluation, with 15% of the data from the training period (T) reserved as an in-distribution test set and 100% of the data from the distinct period (T') used as the OOD test set.

A.10 Related Work

Our work intersects several key areas of research in machine learning, natural language processing, and human-computer interaction. This section reviews relevant literature in modeling LE data, generalization techniques, and the application of LLMs to sequential and structured data, including human-centric applications.

Modeling LE Data Modeling complex, multi-modal LE data is a critical area for diagnostic and prognostic applications in diverse domains, including behavioral and physical health (Nemati et al., 2022; Rabbi et al., 2019; Bae et al., 2017; Xu et al., 2021b), mental health (Wang et al., 2018; Xu et al., 2021a, 2019; Chikersal et al., 2021; Wahle et al., 2016; Farhan et al., 2016; Canzian and Musolesi, 2015; Wang et al., 2022b; Xu et al., 2023a), and education (Wang et al., 2016; Li et al., 2020). Traditional machine learning and deep learning approaches applied to this data, such as time series models or methods based on hand-engineered features, exhibit critical limitations. They often prioritize performance on in-distribution data and struggle significantly with generalizability across datasets exhibiting domain shifts, a challenge notably highlighted by the GLOBEM benchmark (Xu et al., 2023b). Furthermore, they often lack adequate exploration of missing data impact (Xu et al., 2021a; Arnold and Pistilli, 2012) and may not fully capture the complex co-occurrence and relational structure across multi-dimensional LE features (Xu et al., 2019). Training deep neural models on typically limited LE datasets also presents significant challenges (Xu et al., 2023a).

More recently, the potential of LLMs has been explored specifically for LE data forecasting and prediction. Kim et al. (Kim et al., 2024) investigate the capacity of LLMs, using prompting and fine-tuning techniques on multiple health datasets including GLOBEM, to make inferences for various health prediction tasks from wearable sensor data combined with contextual information. While demonstrating promising in-distribution performance and the benefits of context enhancement, their work primarily focuses on within-dataset evaluation and does not extensively study generalizability across datasets or time periods. In parallel, Hayat et al. (Hayat et al., 2024a,b) explore LLM-based LE data forecasting using the MFAFY dataset and diverse LLM architectures. However, consistent with Kim et al., their evaluation focuses on within-dataset performance rather than extensive study of cross-dataset or cross-temporal generalizability. Similarly, Thach et al. (Thach et al., 2025) propose MuHBoost, a multi-label boosting method leveraging LLMs in a zero-shot fashion for predicting multiple health and well-being outcomes using ubiquitous health data, including datasets like GLOBEM and MFAFY. Their work addresses aspects like feature types and missing data, but their evaluation does not specifically investigate the generalizability of the zero-shot LLM approach across different datasets or time periods with domain shifts. While these recent LLM-based studies demonstrate the growing interest in applying foundation models to LE data, they reveal a critical unmet need for methods specifically designed and evaluated for robust *cross-dataset generalizability* under domain shifts, which is a central focus of our ConText-LE framework.

Generalization in Machine Learning Domain adaptation (Pan and Yang, 2010) and domain generalization (Zhou et al., 2022) are key areas in machine learning aiming to improve model performance on target distributions different from the training distribution. While techniques like invariant representation learning, meta-learning, and data augmentation have been explored, their success in complex longitudinal human behavioral data, characterized by multifaceted and often subtle shifts across cohorts and contexts, has been limited (Xu et al., 2023a). In NLP, approaches to improve cross-domain generalization include continued pre-training on domain-specific data (Gururangan et al., 2020), domain-adaptive fine-tuning (Howard and

Ruder, 2018), and prompt-based adaptation (Lu et al., 2022). Our work builds on these insights but focuses specifically on the unique challenges of generalizing across LE data distributions using LLMs as the foundation.

Contextual Representation Learning in NLP The evolution of contextual representation learning in NLP provides important foundations for our work. Early word embedding approaches like word2vec (Mikolov et al., 2013) offered static representations of words, while later models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) revolutionized NLP by introducing dynamic, contextualized representations that capture how a word's meaning changes based on its surrounding context. Recent research has explored how these contextual representation capabilities extend to more complex semantic structures, including frame semantics (Baker et al., 1998) and narrative comprehension (Sap et al., 2019; Liu et al., 2024). Our ConText-LE framework leverages these advances by treating multi-dimensional LE data as a complex semantic structure requiring contextual interpretation. The Meta-Narrative approach specifically draws inspiration from how contextualized models integrate local features into a coherent global representation, addressing the need for both local feature analysis and global contextual synthesis when interpreting complex human behavior patterns.

Prompting Strategies and Reasoning in LLMs Recent advances in prompting strategies have significantly enhanced LLMs' reasoning capabilities. Chain-of-thought prompting (Wei et al., 2023) and similar approaches that break down complex reasoning into intermediate steps have shown remarkable improvements on tasks requiring multi-step inference. Zero-shot reasoning techniques (Kojima et al., 2023) further demonstrate that well-structured prompts can elicit sophisticated reasoning abilities from LLMs without task-specific examples. Our two-stage prompting approach for generating Meta-Narratives builds on these insights, structuring the analysis process into sequential steps of feature analysis followed by contextual synthesis. This approach parallels how humans process complex data—first analyzing individual components before integrating them into a cohesive interpretation—and leverages LLMs' pre-trained understanding of how elements gain meaning through their relationships with other elements. The Prospective Narrative Generation formulation similarly builds on findings that generative formulations often allow LLMs to express complex reasoning more effectively than discriminative ones (Wei et al., 2023; Kojima et al., 2023).

Large Language Models for Sequential and Structured Data LLMs have shown remarkable capabilities not only in natural language processing but also in processing and reasoning about other data modalities when appropriately structured. Approaches for general time series forecasting using LLMs often involve adapting time series data into a format suitable for LLM inputs, such as serialization into sequences of tokens or explicit textual descriptions, followed by fine-tuning or prompting (Sun et al., 2023; Jin et al., 2023; Chang et al., 2023; Gruver et al., 2023; Zhou et al., 2023; Cao et al., 2023; Xue and Salim, 2023; Liu et al., 2023). These methods demonstrate LLMs' potential to capture temporal dependencies and patterns, although challenges remain, particularly with handling the multi-dimensional nature of data and processing long sequences (Liu et al., 2024).

In parallel, LLMs have been applied to human-centric data, leveraging pre-trained knowledge for tasks like health prediction based on textual health records or summarized sensor data (Kim et al., 2024). Most approaches focus on simple encoding strategies like direct verbalization or statistical summarization, while our work explores more sophisticated narrative-based representations. The narrative format aligns with recent findings showing that LLMs perform better when information is presented in coherent, semantically rich formats that leverage their pre-trained understanding of contextual relationships (Wang et al., 2022a; Shwartz et al., 2020). Our ConText-LE framework extends this line of research by developing a specific, structured textual encoding strategy to represent complex, multi-dimensional LE data as a coherent narrative, allowing us to leverage the powerful contextual understanding capabilities of LLMs while preserving the rich semantic relationships between features that might be lost in simpler encoding approaches.

Multimodal Learning for Human Data Multimodal learning, which combines information from different data types or modalities, is increasingly explored for understanding complex human behavior. While some recent work explores multimodal representations for time series or human data by converting them into visual formats and leveraging vision-language models (VLMs) (Zhong et al., 2025), our ConText-LE framework explores an alternative multimodal perspective. By translating multi-dimensional LE data into a *textual* modality, ConText-LE creates a novel cross-modal learning problem where structured behavioral data in one modality is represented and processed using models designed for another (language). This approach aligns with recent work on cross-modal transfer learning (Artetxe et al., 2020) and allows us to investigate the benefits of leveraging the rich semantic space and generalizable patterns learned by LLMs on massive text corpora, applied to the distinct domain of human behavioral sequences.

In summary, while existing work has explored modeling LE data and applying LLMs to time series and human data, achieving robust *cross-dataset generalization* remains a significant challenge, particularly for complex LE data with its inherent multi-dimensionality and domain shifts. Our ConText-LE framework addresses this gap by proposing a novel approach that leverages the contextual representation capabilities of LLMs through a semantically rich narrative representation of multi-dimensional LE sequences, explicitly focusing on improving generalizability across different data distributions.

A.11 Bidirectional Generalization Results

In the main paper, we presented results for the $T \rightarrow T'$ generalization direction, where models were trained on data from the source period (T) and evaluated on data from the target period (T'). In this appendix, we present the complete results for the reverse direction ($T' \rightarrow T$), where models are trained on data from the target period (T') and evaluated on data from the source period (T).

This bidirectional evaluation is crucial for understanding the robustness and symmetry of generalization capabilities. If a method performs well in both directions, it suggests that the approach captures fundamental patterns that are consistent across different contexts, rather than simply exploiting biases specific to a particular generalization direction.

GLOBEM $T' \rightarrow T$ Results Table 4 presents the $T' \rightarrow T$ generalization results for the GLOBEM mental health forecasting task (Year 3&4 → Year 1&2).

For GLOBEM, the $T' \rightarrow T$ results demonstrate consistent superiority of the Meta-Narrative approach across both output formulations. With Binary Classification, Meta-Narrative achieves the highest OOD performance (55.08% accuracy, 50.30% F1), though the margin over other approaches is relatively modest (1.80-2.25% accuracy improvement). Notably, while Meta-Narrative maintains the best overall performance, the precision scores are more competitive across input strategies, with Natural Language String achieving 51.35% precision versus Meta-Narrative's 51.32%.

With Prospective Narrative Generation, the advantages become more pronounced. Meta-Narrative achieves 68.75% OOD accuracy and 66.43% F1, representing a substantial 13.67% absolute accuracy improvement over the same approach with Binary Classification. Natural Language String Encoding shows particularly strong performance in this setting (66.12% accuracy, 66.23% F1), demonstrating that narrative formulations can enhance even simpler representations. The consistent superiority of Prospective Narrative Generation across all input strategies confirms that generative formulations better leverage LLMs' contextual understanding capabilities.

LifeSnaps $T' \rightarrow T$ Results Table 5 presents the $T' \rightarrow T$ generalization results for the LifeSnaps anxiety forecasting task (Last 2 Months → First 2 Months).

The LifeSnaps $T' \rightarrow T$ results reveal striking patterns that emphasize the importance of appropriate representation strategies. With Binary Classification, Statistical Summary encoding demonstrates catastrophic failure on in-distribution data (28.57% F1), highlighting its inability to capture meaningful patterns in the LifeSnaps dataset's specific structure. In contrast, Meta-Narrative achieves robust performance (70.00% ID accuracy, 56.36% OOD accuracy), maintaining the smallest ID-OOD performance gap among all approaches.

Prospective Narrative Generation dramatically transforms the performance landscape. Meta-Narrative achieves exceptional results with perfect balanced performance on ID data (80.00% across all metrics) and strong OOD generalization (71.43% accuracy, 69.23% F1). The 15.07% absolute improvement in OOD accuracy over Binary Classification represents the largest single improvement observed across all datasets and directions. Natural Language String Encoding also benefits substantially from narrative generation, improving

Table 4: **GLOBEM** $T' \rightarrow T$ Generalization Results (Year 3&4 → Year 1&2). Comparison of textual input representation strategies with different output formulations.

	In-Distribution (ID) (Year 3&4 Test)				Out-of-Distribution (OOD) (Year 1&2 Test)			
Input Strategy	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
Output Formulation: Binary Classification								
Complete Sequence	64.14	64.44	58.78	61.48	54.22	52.59	46.73	49.53
Statistical Summary	62.50	62.94	59.60	61.22	52.83	43.87	**51.03**	47.18
Natural Language String	65.79	64.29	57.86	60.90	53.28	**51.35**	46.53	48.82
Meta-Narrative (ours)	**67.43**	**69.23**	**60.40**	**64.52**	**55.08**	51.32	49.32	**50.30**
Output Formulation: Prospective Narrative Generation								
Complete Sequence	68.42	68.38	57.55	62.50	63.16	64.29	59.21	61.64
Statistical Summary	67.11	70.15	61.04	65.28	59.21	65.52	47.50	55.07
Natural Language String	70.39	**70.63**	**62.68**	**66.42**	66.12	69.66	**63.12**	66.23
Meta-Narrative (ours)	**71.71**	69.52	57.48	62.93	**68.75**	**70.15**	63.09	**66.43**

Table 5: **LifeSnaps** $T' \rightarrow T$ Generalization Results (Last 2 Months → First 2 Months). Comparison of textual input representation strategies with different output formulations.

	In-Distribution (ID) (Last 2 Months Test)				Out-of-Distribution (OOD) (First 2 Months Test)			
Input Strategy	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
Output Formulation: Binary Classification								
Complete Sequence	50.00	57.14	66.67	61.54	49.11	54.24	51.61	52.89
Statistical Summary	50.00	25.00	33.33	28.57	46.43	53.33	50.00	51.61
Natural Language String	**80.00**	**100.00**	**66.67**	**80.00**	52.68	50.00	66.04	56.91
Meta-Narrative (ours)	70.00	80.00	**66.67**	72.73	**56.36**	**55.22**	**67.27**	**60.66**
Output Formulation: Prospective Narrative Generation								
Complete Sequence	60.00	57.14	80.00	66.67	62.50	56.60	61.22	58.82
Statistical Summary	50.00	60.00	50.00	54.55	58.04	61.54	42.86	50.53
Natural Language String	60.00	50.00	75.00	60.00	68.75	70.00	63.64	66.67
Meta-Narrative (ours)	**80.00**	**80.00**	**80.00**	**80.00**	**71.43**	**70.59**	**67.92**	**69.23**

from 52.68% to 68.75% OOD accuracy, demonstrating the broader applicability of generative formulations beyond the Meta-Narrative approach.

MFAFY $T' \rightarrow T$ Results Table 6 presents the $T' \rightarrow T$ generalization results for the MFAFY educational engagement forecasting task (Year 2 → Year 1).

The MFAFY $T' \rightarrow T$ results exhibit interesting asymmetries compared to the forward direction. With Binary Classification, Meta-Narrative achieves the strongest OOD performance (68.20% accuracy, 65.60% F1), notably outperforming the forward direction results (60.86% accuracy, 64.96% F1). This 7.34% accuracy improvement suggests that models trained on the more constrained Year 2 data (one semester) may learn more transferable patterns than those trained on the longer Year 1 period (two semesters).

With Prospective Narrative Generation, Meta-Narrative maintains its leadership (70.49% accuracy, 64.00% F1), though Natural Language String Encoding shows competitive performance (68.85% accuracy, 68.33% F1). A notable observation is that the F1 scores remain remarkably consistent across directions for Meta-Narrative (64.14% vs. 64.00%), indicating stable precision-recall balance despite different training contexts. The consistent strong performance across both directions reinforces that Meta-Narrative representations capture domain-invariant educational engagement patterns.

Discussion of Bidirectional Generalization The bidirectional generalization results provide compelling evidence for the robustness of the ConText-LE framework. Our analysis reveals several key insights:

Consistent Meta-Narrative Superiority: Across all datasets and directions, Meta-Narrative input consistently achieves the highest OOD performance, with improvements ranging from 1.80% (GLOBEM Binary) to 15.07% (LifeSnaps Narrative) in absolute accuracy. The approach demonstrates particular strength in challenging scenarios where other methods fail completely (e.g., Statistical Summary on LifeSnaps).

Asymmetric Generalization Patterns: While generalization improvements from narrative ap-

Table 6: **MFAFY** $T' \rightarrow T$ Generalization Results (Year 2 → Year 1). Comparison of textual input representation strategies with different output formulations.

	In-Distribution (ID) (Year 2 Test)				**Out-of-Distribution (OOD) (Year 1 Test)**			
Input Strategy	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
Output Formulation: Binary Classification								
Complete Sequence	60.38	45.58	57.48	51.16	61.48	57.75	58.78	58.26
Statistical Summary	54.72	39.13	47.37	42.86	57.54	48.59	54.98	51.59
Natural Language String	56.60	47.37	40.91	43.90	64.75	59.62	58.49	59.05
Meta-Narrative (ours)	**62.26**	**50.00**	**60.00**	**54.55**	**68.20**	**68.77**	**62.71**	**65.60**
Output Formulation: Prospective Narrative Generation								
Complete Sequence	67.92	61.11	52.38	56.41	66.39	66.07	62.71	64.35
Statistical Summary	66.04	52.38	57.89	55.00	66.72	68.50	49.46	57.44
Natural Language String	66.04	52.49	47.37	50.00	68.85	71.93	65.08	68.33
Meta-Narrative (ours)	**71.70**	**63.16**	**60.00**	**61.54**	**70.49**	**72.73**	**57.14**	**64.00**

proaches are consistent, the magnitude varies significantly by dataset and direction. MFAFY shows better performance in the $T' \rightarrow T$ direction, potentially due to the temporal structure differences between one-semester and two-semester periods. This asymmetry suggests that training data characteristics significantly influence cross-temporal generalization capabilities.

Robust Narrative Generation Benefits: Prospective Narrative Generation consistently outperforms Binary Classification across all datasets and directions, with improvements ranging from 13.67% (GLOBEM) to 15.07% (LifeSnaps). This systematic advantage validates our hypothesis that generative formulations better align with LLMs' inherent capabilities for contextual understanding and reasoning.

Context-Dependent Strategy Effectiveness: The relative performance of different input strategies varies significantly by dataset context. For instance, Natural Language String Encoding performs competitively with narrative generation on MFAFY (qualitative data) but struggles on LifeSnaps (mixed modal data), suggesting that optimal representation strategies may depend on the underlying data characteristics.

The remarkable consistency of these patterns across bidirectional evaluations demonstrates that ConText-LE improvements stem from capturing fundamental data relationships rather than exploiting direction-specific biases. This bidirectional robustness is crucial for practical deployment, where models must perform reliably across diverse temporal contexts and application scenarios.