A MULTIMODAL APPROACH FOR BUILDING GENERALIZABLE AND RELIABLE MODELS OF LONGITUDINAL EXPERIENTIAL DATA

by

Ahatsham

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Engineering

Under the Supervision of Professor Mohammad Rashedul Hasan

Lincoln, Nebraska

May, 2026

A MULTIMODAL APPROACH FOR BUILDING GENERALIZABLE AND RELIABLE MODELS OF LONGITUDINAL EXPERIENTIAL DATA

Ahatsham, Ph.D.

University of Nebraska, 2026

Adviser: Mohammad Rashedul Hasan

This dissertation studies how to build generalizable and reliable models for longitudinal experiential (LE) data. Such data arises in domains where human behavior, experience, and context evolve over time, including education, behavioral health, and related human-centered settings. These datasets are often heterogeneous, partially observed, temporally structured, and vulnerable to out-of-distribution (OOD) shift, making them difficult to model with conventional machine learning pipelines.

The dissertation develops a multimodal research program that progresses from traditional machine learning and deep learning methods to contextual large language modeling, vision-language modeling (VLM), and finally PRISM, a frozen-backbone multimodal framework designed to improve robustness and generalization. Across this progression, the dissertation shows that effective LE modeling depends critically on representation design, contextual reasoning, missingness-aware learning, structure-preserving multimodal encoding, and training constraints that reduce source-specific shortcut learning.

In this dissertation, reliability means more than achieving high in-distribution accuracy. It means producing predictions that remain stable, semantically coherent, and useful when data are incomplete and when the test distribution differs from the training distribution. The results show a clear progression in OOD

performance across the three main benchmarks. On GLOBEM, the strongest OOD accuracy improves from 51.06% with traditional baselines to 67.40% with text-only LLM modeling, 72.86% with LE-Viz, and 79.93% with PRISM. On LifeSnaps, it rises from 48.44% to 67.19%, 71.88%, and 81.25%, and on MFAFY from 50.57% to 64.86%, 66.57%, and 73.71%. Together, these findings show that reliable LE modeling emerges when semantic context, temporal structure, and complementary modalities are aligned with frozen priors that preserve transfer-relevant behavior under distribution shift.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, **Prof. M. R. Hasan**, for his unwavering support, guidance, and mentorship throughout my Ph.D. journey. He has had a profound influence on my growth, not only as a researcher but also as an individual. His patience, trust, and continuous encouragement gave me the confidence to explore ideas, take risks, and push boundaries. He taught me how to truly appreciate research, not just as an academic pursuit, but as a way to create meaningful real-world impact. I am deeply thankful for his belief in me and for the invaluable lessons that will stay with me throughout my life.

I am sincerely grateful to my collaborators **Prof. Bilal Khan, Prof. Neeta Kantamneni, Prof. Heidi Diefes-Dux, Prof. Logan Perry, and Prof. Grace Panther**. Working with each of you has been a rewarding experience, and your insights, perspectives, and guidance have significantly shaped this work. I truly appreciate your support and the collaborative spirit you brought to this journey.

I would like to thank my undergraduate students **Sharif Akil, Helen Martinez, and Hunter Tridle**. It has been a pleasure working with you, and I am grateful for your dedication, curiosity, and contributions. Your enthusiasm made the research process even more enjoyable and meaningful.

I am incredibly thankful to my lab mates and friends **Jiyoung Lee, Chinh Hoang, Jisu Kim, Xuefei Xu, Sofiya Arora, Ashna Chauhan, Pranjay Joshi, Rao Nargis, Kaleem Syed, and Jahangeer**. Beyond research discussions and collaborations, your friendship, encouragement, and shared experiences made this journey memorable. I truly value the support system we built together.

I would like to extend my sincere appreciation to **Teresa Ryans**, ECE Graduate Program Specialist, for her constant support and guidance throughout my time in

the program. Her kindness, responsiveness, and willingness to help at every step made a significant difference. Many aspects of this journey would not have been possible without her support.

I am also deeply grateful to my supervisory committee **Prof. Hamid Sharif-Kashani, Prof. Benjamin Riggan, and Prof. Siamak Nejati**. Thank you for your insightful feedback, thoughtful suggestions, and encouragement. Your guidance helped refine my work and broaden my perspective in meaningful ways.

This dissertation was supported by a standard grant from the U.S. National Science Foundation (NSF DUE 2142558).

Finally, I would like to express my heartfelt gratitude to my family and friends. Their unwavering support, patience, and belief in me have been my greatest strength throughout this journey. During both the challenging and rewarding moments, their presence provided comfort, motivation, and inspiration. This achievement would not have been possible without them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 A Generalization-First View of Longitudinal Experiential data

Human behavior is not static. It unfolds through time, is shaped by context, and is expressed through multiple interdependent channels such as language, actions, routines, physiological rhythms, self-perceptions, and social interactions. In education, a student's performance cannot be understood only through a set of exam scores because it is also shaped by motivation, self-efficacy, engagement, identity, stress, and the broader conditions under which learning occurs. In behavioral health, mental states are not captured only through a single self-report or a single passively sensed variable because changes in sleep, mobility, social interaction, and subjective mood often derive meaning only when interpreted relative to the person's recent trajectory and surrounding context. Similar considerations arise in workplace well-being, clinical monitoring, and human-centered intervention systems more broadly. Across all of these settings, the scientific aim is not merely to classify isolated observations. The aim is to understand and forecast *trajectories* of human experience.

This dissertation uses the term *longitudinal experiential (LE) data* to describe datasets that repeatedly measure aspects of human experience, perception, behav-

ior, and context over time. LE data typically combines heterogeneous observations: self-reported qualitative reflections, numerical measurements, categorical responses, temporal patterns, contextual descriptors, and often incomplete or irregular records. It is therefore richer than conventional tabular data and more semantically charged than standard multivariate sensor streams. Experience sampling and ecological momentary assessment have long demonstrated the value of repeated in-situ measurements for understanding dynamic human states [4–6]. More recent work on personal sensing has further shown that real-world trajectories of behavior can inform personalized modeling of mental health and well-being [7]. In education, longitudinal behavioral and non-cognitive measures provide insights that complement conventional cognitive assessments and can support early intervention [8–10]. In other words, LE data is important precisely because it captures what one-shot measurements cannot: change, progression, persistence, irregularity, and the evolving relationship between context and outcome.

Yet this same richness makes LE data unusually difficult to model. The problem is not simply that LE datasets are temporally structured. Many machine learning (ML) and deep learning (DL) methods already exist for sequential or time-series prediction. The deeper problem is that LE data is *situated*. The meaning of a pattern is often dependent on local context, individual history, and broader environmental conditions. The same observed change can correspond to different latent states in different contexts. Moreover, LE data is frequently expensive to collect, meaning that sample sizes are often modest. Missingness is common often around 30–70% in LE datasets, and can itself be behaviorally meaningful rather than incidental [11–15]. The data may mix qualitative and quantitative signals whose predictive value emerges only when they are interpreted jointly. Finally, the deployment setting is almost always different from the training setting: different

cohorts, different semesters, different institutions, different sensing conditions, different populations, or different time periods. The result is that the decisive challenge is not only predictive performance on a held-out split from the same distribution. It is *generalization under shift*.

This dissertation is titled *A Multimodal Approach for Building Generalizable and Reliable Models of LE data*. Every word in that title reflects a central claim.

Throughout the dissertation, LE models are developed for multiple downstream scenarios, with forecasting as the primary task used to evaluate generalization and reliability.

First, the key criterion is not raw performance alone (i.e., within-distribution benchmark accuracy), but *generalizability*. A model that fits one cohort, one semester, or one data-collection regime but fails when the environment changes has limited scientific value and even less practical value in intervention systems. Cross-distribution robustness is therefore not an auxiliary property but the core property that determines whether LE modeling is meaningful outside a benchmark setting [3, 16–18].

Second, the central challenge is representational. LE data is not naturally reducible to a single modality. It contains numbers, language, context, time, and often structured absence. If it is flattened aggressively into a feature vector, temporal and relational structure may be lost. If it is modeled only numerically, semantic richness is suppressed. If it is serialized only as text, the two-dimensional organization of features through time may collapse into a one-dimensional token stream. This dissertation argues that stronger LE models emerge when representation design explicitly preserves complementary aspects of the underlying trajectory rather than forcing all information into one reduced single-view form [3, 19–22].

Third, the central challenge is also one of *reliability*. In human-centered

domains, an accurate model that is systematically overconfident, brittle under shift, or blind to its own uncertainty is not sufficiently trustworthy for real-world use. Reliable LE systems must therefore be evaluated not only by accuracy or F1, but also by calibration, robustness to stress conditions, and stability under domain shift [23–25].

The research arc of this dissertation develops these claims progressively. It begins by examining why traditional ML and DL methods struggle when asked to forecast outcomes from complex LE trajectories. It then turns to large language models (LLMs), motivated by their ability to interpret contextualized information and leverage rich pretrained knowledge in small-data settings. The next step shows that textualization and narrative reasoning improve alignment with the semantics of LE data, but that text alone still leaves important structure unresolved, especially when missingness and feature–time organization matter. This motivates a multimodal turn in which visual encoding complements language-based representations by preserving structural relationships that plain serialization tends to erase. The dissertation then culminates in a frozen-prior multimodal framework, PRISM (Prospective Reasoning through Integrated Spectral-temporal Multimodal Learning), which argues that robust generalization requires not only better representations but also stronger constraints on learning so that the model is discouraged from fitting distribution-specific shortcuts and is instead pulled toward semantically coherent, transferable structure [1–3, 20–22, 26, 27].

Accordingly, the introduction has four purposes. It first defines LE data as a formal modeling problem and clarifies how it differs from more familiar supervised learning settings. It then positions the methodological transitions of this dissertation—from classical ML/DL to LLMs, from text-only to multimodal learning, and from unconstrained fine-tuning to frozen-prior training—as one

continuous scientific argument rather than a sequence of disconnected papers. It next situates the dissertation within the relevant literature on educational data mining, behavioral sensing, language modeling, multimodal learning, missing-data modeling, and domain generalization. Finally, it states the dissertation contribution, the research questions, and the chapter-level organization of the dissertation.

1.2 LE data: Definitions, Structure, and Scientific Value

The phrase *LE data* is used in this dissertation in a deliberately broad but operational sense. It refers to repeated observations of an individual or unit over time where at least part of the signal captures lived experience, perception, subjective state, or behavior in context. The observations can be self-reported, passively sensed, administratively recorded, or derived from interactions with systems. What distinguishes LE data from ordinary repeated measures is that interpretation depends on trajectory, circumstance, and cohort-level context, not just on the absolute value of isolated variables. Put differently, standard time-series analysis asks how one variable evolves over time, while longitudinal modeling asks how many variables evolve across people, how those evolutions differ, and what explains those differences.

Formally, let the trajectory for individual i across T time steps be written as

$$X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,T}\}, \tag{1.1}$$

where each $x_{i,t}$ is not a scalar but a structured observation containing multiple

views of the person at time t. A convenient abstraction is

$$x_{i,t} = \left(x_{i,t}^{(n)}, x_{i,t}^{(o)}, x_{i,t}^{(c)}, x_{i,t}^{(q)}, m_{i,t}, \delta_{i,t}\right), \tag{1.2}$$

where $x_{i,t}^{(n)}$ denotes numeric variables, $x_{i,t}^{(o)}$ ordinal variables, $x_{i,t}^{(c)}$ categorical variables, $x_{i,t}^{(q)}$ qualitative or textual content, $m_{i,t}$ a missingness mask, and $\delta_{i,t}$ a vector of time-gap indicators capturing irregular observation intervals. A forecasting objective then seeks to infer a future state from a historical window:

$$\hat{y}_{i,t+h} = f(x_{i,t-w+1}, \ldots, x_{i,t}), \tag{1.3}$$

where w is the lookback window and h the prediction horizon.

This notation highlights a key point: in LE settings, the object to be modeled is not merely a matrix of numbers. It is a multi-view longitudinal object in which semantics, temporality, context, and incompleteness may all be informative. To illustrate, consider three representative settings that recur throughout this dissertation.

In **education**, LE data may include repeated reflections about lecture engagement, confidence, performance self-evaluation, science identity, and other non-cognitive constructs, alongside assessments and background factors. Such data captures patterns that are predictive of future academic performance but are not reducible to test scores alone [8–10]. A student who appears average on early assessments may nevertheless be on a trajectory of disengagement that becomes visible only through repeated qualitative or non-cognitive measures. Conversely, a student with low early scores but improving engagement and self-efficacy may be on a rising trajectory that a purely cognitive model would miss. The temporal

evolution of non-cognitive features therefore matters as much as the final values themselves.

In **mental and behavioral health**, LE data often combines self-reports and passive sensing. Mood, sleep, mobility, phone usage, activity patterns, and social interaction signals may be measured repeatedly to support prediction of mental-health or well-being outcomes [7, 16, 28]. Here too, absolute values are rarely self-interpreting. Reduced activity might signal depressive worsening, but it might also reflect contextual pressures such as an exam period, illness recovery, work overload, or temporary lifestyle changes. Meaning is inseparable from context.

In **workplace or daily-life well-being**, longitudinal observations may capture stress, workload, exhaustion, routine changes, and subjective well-being in settings where the dynamics of recovery, pressure, and adaptation are central [7, 29]. Again, the predictive signal lies not in any single variable but in the relationships among variables over time and the circumstances under which they evolve.

These examples show why LE data demands more than generic forecasting machinery. It possesses at least five properties that are central to this dissertation.

1.2.1 Context-dependence

A recurring claim throughout this dissertation is that experiential observations are context-dependent rather than fixed in meaning. Let $\mathcal{C}_{i,t}$ denote the local and global context surrounding $x_{i,t}$. Then the meaning of an observation is better described as

$$\text{Meaning}(x_{i,t}) = g(x_{i,t}, \mathcal{C}_{i,t}), \tag{1.4}$$

rather than as a function of $x_{i,t}$ alone. This deceptively simple formulation has deep implications. It means that a model that treats features as context-independent

carriers of meaning may fail precisely where LE modeling is most useful. A decrease in social interaction, for example, can reflect isolation in one setting and focused work in another. A rise in self-reported concern may indicate vulnerability in one individual but productive metacognitive awareness in another. In educational trajectories, confidence, satisfaction, and worry can be predictive only when interpreted with respect to the learner's prior path and surrounding circumstances.

1.2.2 Heterogeneity

LE data is not homogeneous. Variables differ in scale, modality, interpretability, and frequency. An abstract state-space view is

$$x_{i,t} \in \mathbb{R}^{d_n} \times \mathbb{Z}^{d_o} \times \mathcal{C}^{d_c} \times \mathcal{T}^{d_q}, \tag{1.5}$$

where each component may obey different statistical and semantic regularities. This heterogeneity is not merely a technical nuisance. It is often the source of predictive strength because different views capture complementary aspects of the same latent process. A student's self-efficacy, for instance, may interact with assessment results differently from how phone usage interacts with sleep in a behavioral-health dataset. The central design question becomes how to preserve these heterogeneous signals without destroying the relationships that make them meaningful.

1.2.3 Temporal dependency

LE data is not i.i.d. across samples or across time. It embodies persistence, progression, abrupt change, cyclical patterns, and lagged interactions. Two trajectories with the same endpoint can reflect very different processes if one is stable and

the other is volatile. Early forecasting tasks make this even harder because they ask the model to act before the trajectory has fully unfolded. This is one reason why LE forecasting is often closer to interpretation under partial information than to standard sequence completion.

1.2.4 Missingness and irregularity

Missing data is the norm rather than the exception in human-centered longitudinal studies, with missingness rates often reported around 30–70% in LE settings [11–15]. Participants skip prompts, sensors fail, institutions change instruments, and observation schedules become irregular. Rubin's classic taxonomy distinguishes missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [30]. In LE settings, MNAR is especially important because the fact that something is missing may itself carry behavioral signal. A skipped reflection, a missing questionnaire, or a gap in sensing may reflect disengagement, burden, emotional avoidance, routine disruption, device non-use, or other meaningful causes. If missingness is treated only as noise to be numerically patched, this signal may be lost or distorted. Later chapters of this dissertation show that missingness should often be modeled as semantics rather than solely as absence [1, 31, 32].

1.2.5 Multimodal structure

Even when the raw data is stored in a table, its conceptual structure is often multi-modal. Feature values form temporal patterns, qualitative responses form narratives, missingness forms masks, and interactions among these elements may be easier to interpret visually or linguistically than numerically. This is why the representation question is so central. The fundamental object is not just a table. It

is a structured, heterogeneous, multi-view trajectory.

The scientific value of LE data follows directly from these properties. By preserving temporal and contextual information, LE data makes it possible to model outcomes in a way that respects how human states evolve in the real world. This is particularly important in intervention settings, where the relevant question is not "What label best describes this sample?" but "What is happening to this person, how is it changing, and what is likely to happen next?" The goal is not merely prediction, but interpretable and actionable prediction.

1.3 The Core Difficulty: Distribution Shift, Data Scarcity, and Reliability

Because LE data is rich, nuanced, and behaviorally complex, the practical challenge is severe: models trained on one distribution often degrade sharply on another. This dissertation takes a *generalization-first* view because the deployment problem in LE modeling is almost always a domain-shift problem.

Let $(X, Y) \sim P_{\text{train}}$ denote the source distribution used for model fitting and $(X, Y) \sim P_{\text{test}}$ denote the target distribution encountered at evaluation or deployment. In LE applications, it is common that

$$P_{\text{train}}(X, Y) \neq P_{\text{test}}(X, Y). \tag{1.6}$$

The shift may arise from changes in population composition, collection period, seasonality, sensing conditions, institutional context, or intervention regime. For educational data, one semester differs from another; for behavioral sensing, one cohort or year differs from another; for qualitative self-reports, social and contextual

changes alter how participants express themselves.

This problem has been formalized in machine learning as domain adaptation or domain generalization [17, 18]. While the formal literature is broad, one useful intuition comes from generalization bounds that relate target-domain risk to source-domain risk plus a divergence term between source and target distributions. If a model fits the source domain by exploiting correlations that are specific to that domain, then source performance can remain high even while target performance collapses. In human-centered data, this danger is amplified because many surface regularities are context-bound. A model might learn semester-specific, cohort-specific, or platform-specific shortcuts that correlate with labels in training but fail to transfer.

The GLOBEM (Generalization of LOngitudinal BEhavior Modeling) benchmark, a widely used LE benchmark for cross-distribution mental-health forecasting, makes this challenge explicit [16]. Rather than rewarding models that perform well on random train–test splits, GLOBEM emphasizes cross-dataset and cross-temporal generalization in longitudinal behavioral modeling. The benchmark demonstrates that models that appear strong under conventional evaluation can approach chance-level performance under genuine distribution shift. This result is not an anomaly. It shows that cross-distribution robustness is where human-centered longitudinal modeling is hardest and most necessary. The dissertation's subsequent frameworks—ConText-LE (narrative LLM forecasting), LE-Viz (visual-text multimodal forecasting), and PRISM (frozen constrained multimodal learning) were developed in part to address this gap [3, 21, 22].

A second challenge is data scarcity. Many LE datasets are small because collecting repeated human measurements is expensive and burdensome. Studies may involve dozens to hundreds of participants rather than millions of sequences.

Labels may be sparse, and the number of effective training trajectories can shrink further when tasks are defined using sliding windows, outcome filtering, or domain-specific constraints. In such settings,

$$N \ll \text{complexity}(X), \tag{1.7}$$

where N is the number of training trajectories and $\text{complexity}(X)$ reflects the dimensionality, modality, and temporal structure of the data. The challenge is not only classical overfitting. It is that richly parameterized models can fit superficial patterns that look predictive in the source distribution but do not correspond to transferable behavioral structure.

A third challenge is reliability. Suppose a model outputs class probabilities $\hat{p}(y \mid X)$. In many applications, a high score is interpreted as a high-confidence forecast. But modern neural networks are often poorly calibrated, meaning that confidence does not reliably correspond to empirical correctness [23]. In LE settings, this matters because overconfident mistakes can lead to poor interventions or misplaced trust. Reliability therefore requires more than average performance. It requires calibration, robustness under perturbation, and ideally some sensitivity to domain shift itself. This dissertation treats reliability as integral to the modeling problem rather than as a downstream add-on [24, 25].

Collectively, these concerns motivate a view of LE modeling that differs from standard supervised learning. The objective is not simply to maximize predictive accuracy on i.i.d. held-out data. The objective is to construct representations and learning procedures that preserve the aspects of trajectories most likely to remain meaningful when context changes. This is where the dissertation's central argument begins: *LE generalization depends on representation quality and learning*

constraints at least as much as on model size or architectural sophistication.

1.4 Why Traditional Machine Learning and Deep Learning Are Necessary but Insufficient

The first stage of this research program examined traditional machine learning and deep learning approaches to LE forecasting. These methods remain important for at least three reasons. They offer strong baselines, they clarify what is difficult about the problem, and they provide a point of comparison for later shifts to LLMs, multimodality, and frozen priors.

In educational forecasting and related domains, a common pipeline begins by engineering summary features from a trajectory and then fitting a predictive model such as logistic regression, support vector machines, random forests, gradient-boosted trees, multilayer perceptrons, recurrent networks, or temporal convolutional networks [8–10]. When enough labeled data is available and the evaluation split is close to i.i.d., these methods can be effective. They are computationally efficient, interpretable in limited settings, and often competitive on moderately sized benchmarks.

However, several assumptions underlying such approaches become fragile in LE settings.

1.4.1 Flattening and aggregation can destroy structure

A common preprocessing step is to convert a structured feature–time trajectory into a fixed vector:

$$X_i \in \mathbb{R}^{T \times F} \quad \longrightarrow \quad \mathrm{vec}(X_i) \in \mathbb{R}^{TF}, \tag{1.8}$$

or to aggregate across time using means, slopes, variances, and other summary statistics. While convenient, this step can erase precisely the structure that makes LE data informative. Temporal adjacency, localized transitions, co-occurring feature changes, and irregular but meaningful patterns may become hard to recover once the trajectory is flattened or averaged. Two trajectories with the same mean can still reflect opposite underlying dynamics. Flattening treats temporal organization as incidental when, in LE data, it is often the source of meaning.

1.4.2 Numerization can suppress semantics

Many LE variables are not naturally numeric, even if they can be numerically encoded. Qualitative reflections, engagement descriptions, self-assessments, and contextual notes can be coerced into integers or one-hot vectors, but that encoding discards the semantic structure that a human would use to interpret them. This is especially problematic for longitudinal experiential responses, where subtle differences in language or context may be the main predictive signal. Traditional models may only see sparse categories or ordinal values where richer meaning exists.

1.4.3 Small-data deep learning remains unstable

Sequence models such as long short-term memory networks (LSTMs) and Transformer forecasters are attractive for temporal prediction [33, 34], but they often assume data regimes and problem structures different from those found in LE applications. When training data is limited, qualitative, heterogeneous, and incomplete, the benefits of architectural capacity do not necessarily materialize. Instead, such models may memorize idiosyncrasies of the source distribution or fail to exploit the contextual information needed for interpretation. Moreover,

standard sequence architectures are usually optimized for numeric forecasting rather than for semantically rich reasoning.

This limitation was evident in early forecasting experiments that motivated the dissertation. In limited-data educational settings, training neural models from scratch on cognitive trajectories was insufficient for high-quality early prediction, particularly when only a small portion of the semester was observed. This led to the exploration of transfer learning through pretrained language models [26, 27]. Similarly, later work on qualitative non-cognitive longitudinal data showed that direct application of numeric baselines such as Random Forests [35] and LSTMs performed poorly relative to language-driven approaches, suggesting that the problem was not only one of sequence learning but of semantic alignment between the data and the model [2].

1.4.4 in-distribution (ID) success can obscure OOD brittleness

Perhaps the most important limitation is that classical pipelines can appear adequate under narrow evaluation but fail under domain shift. If train and test sets are drawn from the same environment, the model may exploit features that correlate with outcomes only within that environment. In LE modeling, this is especially dangerous because many apparent predictors are really proxies for time period, cohort, or context. The GLOBEM benchmark showed that this brittleness is not hypothetical [16]. Later chapters of this dissertation repeatedly return to the same lesson: it is possible to build systems that look competent under standard validation and yet fail to model the underlying process in a transferable way.

For these reasons, traditional ML and DL were essential to the dissertation not because they solved the problem, but because they exposed its true difficulty. They made clear that better generalization would require more than larger re-

current models or better feature engineering. It would require a change in the representational interface between LE trajectories and the model.

1.5 The Language Turn: Why LLMs Became a Natural Interface

The transition to LLMs in this research program was motivated by a representational insight: LE trajectories often contain semantics that are better expressed through language than through raw numeric encoding.

Pretrained language models are powerful partly because they learn distributed representations shaped by vast amounts of text. They can interpret entities relative to context, model long-range dependencies, and leverage prior knowledge about human situations, states, and relationships [34, 36–39]. In small-data domains, this prior knowledge can compensate for the lack of task-specific data if the task can be expressed in a language-compatible form. This suggests a natural strategy for LE data: instead of forcing the model to operate directly on sparse, heterogeneous trajectories, convert the trajectory into a semantically meaningful textual representation and let the model reason over that representation.

Formally, let $\mathcal{V}$ denote a verbalization operator that maps a structured trajectory to text:

$$T_i = \mathcal{V}(X_i). \tag{1.9}$$

A forecasting model can then be learned as

$$\hat{y}_{i,t+h} = f_{\mathrm{LM}}(T_i), \tag{1.10}$$

or more generally as a generation problem in which the model produces a prospective description from which the target label can be extracted. The central question

becomes whether $\mathcal{V}$ can encode the trajectory in a way that preserves the relevant semantics while aligning with the strengths of pretrained language models.

This idea first appeared in the dissertation through work on early forecasting of academic performance [26, 27]. In those studies, students' early-semester trajectories were verbalized and supplemented with distal background factors and proximal non-cognitive variables. Forecasting was reframed as language generation rather than solely as classification. Two themes emerged immediately.

Personalization. Distal factors such as academic meta-information or background conditions provide a personalized context through which current observations can be interpreted. The same current score or same early-semester pattern may imply different future outcomes depending on the student's broader profile.

Contextualization. Proximal non-cognitive and experiential variables provide local context that enriches cognitive measurements. Instead of using assessment scores alone, the representation integrates engagement, perceptions, and other signals that help a language model understand the trajectory as part of a broader learning process.

These early studies [26, 27] showed that language models can be surprisingly effective in limited-data educational forecasting because the verbalized representation turns the problem into a form more compatible with transfer learning. The model benefits not only from sequence modeling capacity, but also from semantic priors embedded in pretraining. A similar insight motivated later work on broader LE data in education, where richer LE trajectories—including qualitative, cognitive, and contextual components—were processed through pre-trained LMs [20]. That work highlighted a broader point: once trajectories are reframed as semantically meaningful descriptions, the language model is no longer just a classifier. It becomes a reasoner over human-centered longitudinal processes.

This shift from numbers to language mattered for two reasons. First, it addressed the mismatch between small data and high-capacity models. Rather than training a large temporal model from scratch, the approach leveraged pretrained linguistic knowledge. Second, it redefined the problem itself. Forecasting was no longer viewed as merely mapping a vector to a label; it became an interpretive task in which future outcomes are inferred from the narrative meaning of past experiences and behaviors.

The dissertation's later chapter on LLM-based approaches extends this idea in two important directions. The first is to treat language not only as an input interface but also as an output interface. In ConText-LE, forecasting is reframed as *Prospective Narrative Generation*, where the model produces a description of a likely future state rather than directly emitting a binary class [3]. The second is to treat narrative construction itself as a representation problem. Not all verbalizations are equally useful. A list-like serialization of raw values is much less aligned with LLM strengths than a semantically coherent summary of temporal patterns and context. This insight leads to the dissertation's notion of *Meta-Narrative* representations, which synthesize trajectories into higher-level descriptions designed to preserve behavioral meaning while remaining language-native.

In this sense, the move to LLMs is best understood as a move toward *semantic interface design*. The core question is not whether a larger model will automatically solve LE forecasting, but how LE trajectories should be represented so that a pretrained language model can deploy its contextual reasoning capacity on them effectively.

1.6 Missingness as Semantics and Forecasting as Narrative Reasoning

The LLM turn made progress, and it also surfaced two problems that traditional pipelines had obscured: missingness and interpretation.

The first problem is that missingness is not a simple nuisance variable. In many LE datasets, observations are missing for reasons connected to the phenomenon of interest. A student may skip a reflection because of disengagement, overload, or avoidance. A participant may stop wearing a device because routines change, because burden increases, or because well-being deteriorates. A missing value is therefore often an incomplete observation of a latent behavioral event rather than a blank space in a numeric table.

Classical imputation methods estimate missing values numerically under statistical assumptions about the data-generating process [30–32]. Such methods are useful, but they are not always conceptually well matched to LE settings, especially under MNAR conditions where the act of missingness itself is informative. This motivated the dissertation's work on **Contextually Relevant Imputation leveraging pre-trained Language Models (CRILM)** [1]. Instead of treating imputation solely as numeric estimation, CRILM uses language models to generate contextually relevant descriptors for missing values, effectively transforming absence into semantically interpretable information that downstream models can use. The key conceptual move is that imputation is treated not merely as data repair but as representation enrichment.

The second problem is that once trajectories are converted into language, the output space need not remain a single label. A direct binary classification objective forces a rich trajectory into a low-bandwidth output. By contrast, narrative gener-

ation permits the model to express richer future-oriented structure. ConText-LE developed this idea through **Prospective Narrative Generation**, arguing that a model may generalize better when it predicts future states by generating semantically grounded narratives instead of directly selecting from a small label set [3]. This aligns forecasting with the strengths of LLMs: contextual interpretation, generative reasoning, and flexible semantic composition.

The shift can be written abstractly as moving from

$$f : X \to Y \tag{1.11}$$

to

$$g : \mathcal{V}(X) \to \mathcal{N}, \qquad h : \mathcal{N} \to Y, \tag{1.12}$$

where $\mathcal{N}$ denotes a narrative space and h extracts or interprets the forecast from the generated narrative. The value of this formulation is not only flexibility. It imposes a semantic bottleneck. A model that must generate a coherent future-oriented narrative may be less able to rely on brittle surface shortcuts than a classifier that only needs to separate labels within the source distribution. ConText-LE empirically supported this view by showing that narrative-based formulations can improve cross-distribution generalization across diverse LE datasets [3].

A related contribution appears in the dissertation's work on forecasting student engagement from qualitative longitudinal data [2]. There, the full pipeline combines LLM-informed imputation, zero-shot feature selection, and fine-tuning on textual non-cognitive responses. The broader message of this work is that when the raw data is qualitative, sparse, and irregular, language models offer a coherent end-to-end framework in which representation, imputation, selection, and prediction all occur within a semantically aligned space.

These developments sharpened the dissertation's central claim. The problem was no longer simply "Can language models do better than numeric baselines?" The deeper question became "How should human-centered trajectories be represented so that missingness, temporal change, and context are treated as meaningful components of the state rather than as noise around a numeric core?"

1.7 Why Text Alone Still Leaves a Structural Deficit

Despite the advantages of verbalization and narrative reasoning, text-only approaches do not fully solve the LE representation problem. This dissertation identifies a structural deficit that emerges when complex temporal data is forced into a one-dimensional token sequence.

Consider an LE trajectory represented as a matrix over features and time:

$$X_i \in \mathbb{R}^{F \times T}. \tag{1.13}$$

In the raw structure, adjacency exists both across neighboring time points for the same feature and across co-occurring features at the same time. These local neighborhoods can be behaviorally meaningful. For example, abrupt changes in two related variables occurring in the same week may signal a transition that is easy to see in a structured matrix or chart. When the same matrix is serialized into text, however, the two-dimensional organization is linearized:

$$\mathcal{S} : \mathbb{R}^{F \times T} \to (w_1, w_2, \ldots, w_L), \tag{1.14}$$

where $w_1, \ldots, w_L$ is a token sequence. The order imposed by serialization necessarily privileges one traversal of the matrix and destroys others. Relationships that

were spatially local can become token-distant. Feature co-occurrence can be separated by lengthy text spans. Temporal trends can become harder to recover when repeated for many variables. In short, serialization provides semantic accessibility but not necessarily structural faithfulness.

This issue becomes more severe as the number of features and time steps grows. Text can summarize or narrate a trajectory, but it can also blur its internal organization. A list of variable values, even when verbalized, may not permit the model to recover the feature–time geometry that was explicit in the original representation. This insight emerged clearly in the lead-up to LE-Viz [21]. The question was no longer whether language helps; it was whether language alone can preserve enough of the trajectory's structure for robust generalization.

The answer proposed by this dissertation is that *text modality alone cannot capture all of this structure*. Language is excellent for capturing meaning, abstraction, and contextual synthesis. It is less naturally suited to preserving the geometric organization of high-dimensional temporal matrices. If semantic and structural information are complementary, then the representation should be complementary as well.

This argument also reframes the role of multimodality. Multimodality is not introduced here as a generic "more modalities are better" claim. It is introduced because different modalities preserve different invariants of the same underlying object. Text preserves interpretive and semantic invariants; visual encodings preserve structural and spatial invariants. The dissertation's multimodal chapters arise directly from this representational diagnosis.

1.8 The Multimodal Turn: Visual Encodings as Structure-Preserving Operators

The shift from text-only modeling to multimodal modeling in this dissertation is motivated by a precise representational need: to preserve structure that textual serialization tends to collapse.

Visual representations are useful because they can keep feature–time relationships spatially organized. A line chart preserves local continuity within a feature across time. A heatmap preserves the matrix layout of multiple features over multiple steps. A chain of aligned temporal panels can preserve progression and local changes. These visual encodings do not merely make the data prettier or more accessible to humans. They change the inductive form of the input presented to the model. A VLM can attend jointly to spatial structure and textual semantics, potentially recovering relationships that neither modality captures as effectively alone [19, 40].

LE-Viz is the dissertation's first explicit multimodal answer to this problem [21]. The framework transforms LE trajectories into complementary textual and visual representations and uses a VLM to generate prospective narratives. The underlying claim is that the visual channel acts as a structure-preserving operator over the feature–time matrix, while the textual channel provides semantic interpretation. Together, they create a representation that is both more informative and better aligned with the model's strengths than either channel alone.

This idea can be formalized abstractly as learning over paired representations

$$z_i = \Phi_{\text{text}}(X_i) \oplus \Phi_{\text{vis}}(X_i), \tag{1.15}$$

where Φ_{text} maps the trajectory to a textual or narrative representation, Φ_{vis} maps it to a visual encoding, and $\oplus$ denotes a fusion operation. The key question is not only whether z_i contains more information, but whether it contains more *transferable* information. LE-Viz argues that spatial organization itself contributes to cross-distribution robustness because it preserves the structural relationships among co-occurring features and nearby time points in a form that a VLM can exploit [21].

This multimodal turn is also significant methodologically. It expands the dissertation's representational argument beyond language. Earlier chapters showed that semantic alignment matters. LE-Viz shows that semantic alignment is necessary but insufficient when structural relationships are central to the task. Multimodal learning therefore becomes a principled design choice rather than a mere scaling choice.

It is important, however, not to overstate what multimodality solves automatically. Adding modalities can increase computational cost, introduce new fusion challenges, and invite overfitting when data is limited. End-to-end adaptation of large VLMs on small LE datasets is especially risky because it expands the trainable hypothesis space. This observation leads directly to the dissertation's final step: multimodal learning under frozen priors and explicit representational constraints.

1.9 Frozen Priors, Constraint-Based Learning, and the Emergence of PRISM

If better representations improve LE generalization, why do strong pretrained models still fail under distribution shift? The culminating argument of this disser-

tation is that introducing complementary representations alone is not enough. The learning dynamics themselves must be constrained so that the model cannot easily drift toward distribution-specific shortcut solutions.

This insight motivates **PRISM**—*Prospective Reasoning through Integrated Spectral-temporal Multimodal Learning*—the final framework of the dissertation [22]. PRISM is built on the claim that behavioral generalization fails not simply because models lack capacity, but because they are insufficiently constrained during adaptation. A standard prediction loss can often be minimized by learning source-specific correlations that do not encode the deeper behavioral structure needed for transfer. Even narrative generation, when fully fine-tuned, may let the meaning of "coherent" adapt toward the source distribution. PRISM addresses this by introducing a frozen language prior as a distribution-invariant coherence constraint.

At a high level, PRISM optimizes

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{coh}}, \tag{1.16}$$

where $\mathcal{L}_{\text{pred}}$ is a forecasting loss and $\mathcal{L}_{\text{coh}}$ is a narrative coherence loss defined with respect to a frozen pretrained language model. The crucial point is that the coherence criterion does not co-adapt with the downstream task. Because the language prior is frozen, the representation is forced to satisfy a stable surface of semantic plausibility rather than one that can shift opportunistically during training. PRISM further decomposes trajectories into complementary streams—temporal measurements, spectral dynamics, and semantic meta-narratives—and fuses them through directed pairwise interactions and gated aggregation [22].

Why might freezing help? One argument comes from capacity control. If a large pretrained backbone is fully fine-tuned in a small-data, high-shift regime,

the effective hypothesis class becomes enormous. Freezing most of the pretrained structure and adapting only limited components reduces the degree of freedom available for memorizing source-domain peculiarities [41, 42]. A second argument comes from invariance: if the model must remain compatible with a fixed pretrained semantic prior, it may be less free to invent representations that predict labels accurately on the source domain but fail to correspond to transferable behavioral semantics. A third argument comes from reliability: constrained adaptation often improves calibration and reduces the overconfident behavior that can arise when a model overfits small source domains [23, 24].

PRISM therefore represents more than another architecture. It is the conceptual endpoint of the dissertation's trajectory. The early chapters show that LE generalization improves when trajectories are expressed in semantically richer representations. The middle chapters show that missingness and narrative output formulation matter. The multimodal chapter shows that structure preservation matters. PRISM unifies these insights into a stronger claim: *generalization in LE modeling emerges when the model is given complementary representations and is constrained by frozen priors that prevent shortcut learning from dominating training*. This is where the dissertation's concern with reliability becomes explicit, because the same constraints that encourage transfer can also improve calibration and reduce brittle overconfidence.

1.10 Literature Positioning: Where This Dissertation Sits in the Broader Research Landscape

The dissertation sits at the intersection of several research literatures that have often developed in partial isolation: longitudinal behavioral modeling, educational

data mining, time-series learning, missing-data modeling, LLMs for structured data, multimodal learning, and domain generalization. Its contribution is not to replace these literatures, but to synthesize and extend them for a specific scientific problem: building LE models that remain useful under real-world shift.

1.10.1 Experience sampling, ecological momentary assessment, and personal sensing

The methodological roots of LE data lie in repeated in-situ measurement. experience sampling methods (ESM) and ecological momentary assessment (EMA) established the value of collecting data in real time or near real time to capture dynamic human states that retrospective measurement can distort [4–6]. Personal sensing and mobile health work later extended this idea by combining subjective reports with passive behavioral streams from ubiquitous devices [7]. These literatures established the scientific importance of LE data but did not by themselves solve the machine learning problem of generalization under shift.

1.10.2 Educational forecasting and behavioral modeling

Educational data mining and learning analytics have shown that both cognitive and non-cognitive variables can support early intervention [8–10]. Much of this work focused on prediction from structured numerical features and often evaluated under within-cohort conditions. The early papers in this dissertation build on this foundation while arguing that small-data educational settings benefit from transfer learning and contextualized representations rather than from purely task-specific numeric models [26, 27].

1.10.3 Missing-data modeling

The statistical literature on missing data has long emphasized that missingness mechanisms matter [30]. More recent machine learning work has introduced temporal imputation models such as GRU-D and BRITS for multivariate time series [31, 32]. These approaches are important, but they generally treat the end goal as estimating plausible numeric values. This dissertation contributes a complementary perspective by showing that in LE settings, especially under MNAR-like conditions, semantically relevant textual descriptors can be a better representational choice for downstream modeling than purely numeric imputation [1].

1.10.4 Foundation models for structured and temporal data

A growing literature explores how LLMs can operate on tables, time series, and other non-linguistic data by converting them to text or aligning them with language-model interfaces. This includes work on table understanding such as TAPAS, as well as broader work on foundation models and transfer learning [37,43]. The dissertation contributes to this literature by focusing on LE trajectories as a particularly challenging class of structured temporal data: context-dependent, missingness-rich, small-data, and shift-prone. It further argues that not all serializations are equally suitable; narrative formulations aligned with the model's contextual reasoning capabilities can yield stronger generalization than direct raw serialization [3].

1.10.5 Multimodal learning and VLMs

Multimodal machine learning provides the conceptual basis for combining complementary representations [19]. Recent VLMs show that pretrained systems can reason jointly over visual and textual information. LE-Viz extends this idea to LE trajectories by using visual encoding not as decoration but as a structure-preserving representation of the feature–time organization that text serialization tends to flatten [21, 40]. This moves the dissertation from language-centered to explicitly multimodal modeling.

1.10.6 Domain generalization and reliability

The formal literature on domain adaptation and domain generalization provides a vocabulary for understanding why models trained on one environment fail on another [17, 18]. Reliability research in machine learning highlights that confidence estimates from modern neural models are often miscalibrated [23–25]. PRISM brings these concerns together by arguing that frozen-prior constraints can improve both transfer and reliability in LE forecasting [22].

Seen in this broader context, the dissertation's position is distinctive in three ways. First, it treats *LE data* as a unified object of study spanning education, mental health, and related domains. Second, it approaches the problem as one of *representation and constrained transfer*, not merely of bigger models. Third, it makes *cross-distribution generalization and reliability* the primary criteria by which methodological advances should be judged.

1.11 Central Claim of the Dissertation

The central claim of this dissertation is the following:

Generalizable and reliable modeling of LE data requires representations that jointly preserve semantic context, temporal organization, and complementary behavioral structure across modalities, together with learning objectives and adaptation strategies that constrain the model toward transferable, semantically coherent reasoning rather than distribution-specific shortcuts.

This dissertation has several implications.

First, it implies that LE modeling is fundamentally a representation problem. Improvements will not come only from increasing model size or trying more architectures on the same flattened inputs. They will come from redesigning how trajectories are expressed to the model.

Second, it implies that language is not merely a convenient interface but a scientifically meaningful representational medium for LE data. Verbalization, contextualization, narrative forecasting, and semantically informed imputation are not cosmetic transformations; they expose behavioral meaning to pretrained models that would otherwise struggle to exploit it.

Third, it implies that text alone is not always sufficient. When the organization of the feature–time matrix carries predictive signal, structural preservation becomes essential, motivating multimodal learning with visual encodings.

Fourth, it implies that reliable transfer requires constraint. Frozen priors, multi-objective learning, and restricted adaptation spaces help prevent models from solving the source task in ways that fail to transfer.

Finally, it implies that the appropriate end point for LE modeling is not merely higher ID accuracy. It is a system that remains useful when the world changes and that knows, as much as possible, when it may be wrong.

1.12 Research Questions and Working Hypotheses

The dissertation is organized around the following research questions.

1. **RQ1.** Why do traditional ML and DL methods struggle to model LE trajectories under small-data conditions, heterogeneity, missingness, and domain shift?
2. **RQ2.** How can LLM-based verbalization and contextualization improve LE forecasting by aligning trajectory representation with pretrained semantic knowledge?
3. **RQ3.** How should missingness be modeled when absence itself may be behaviorally meaningful?
4. **RQ4.** Why is text-only serialization insufficient for some LE tasks, and how can multimodal visual–text representations preserve structure that supports cross-distribution transfer?
5. **RQ5.** What training constraints are needed for multimodal LE models to improve generalization and reliability under distribution shift?

These questions motivate a set of working hypotheses that guide the empirical chapters.

- **H1.** Contextualized and semantically enriched language representations outperform conventional numeric baselines for LE forecasting in limited-data settings.
- **H2.** Treating missing values through context-aware semantic descriptors can improve downstream forecasting, particularly when missingness is not random.

- **H3.** Narrative-based formulations of forecasting improve OOD performance by aligning the task with the generative strengths of LLMs.
- **H4.** Multimodal representations that preserve the feature–time structure visually, while also providing textual semantics, reduce the information loss caused by one-dimensional serialization.
- **H5.** Frozen-prior constraints improve both transfer and reliability by reducing the model's freedom to learn source-specific shortcuts.

1.13 Methodological Principles and Evaluation Philosophy

Generalization and reliability in this dissertation require a clear account of how those properties are evaluated. One of the recurring problems in applied machine learning for human-centered data is that success is often declared on the basis of evaluation protocols that are much easier than the deployment conditions the model will eventually face. Random train–test splits, isolated accuracy numbers, and single-run summaries can make a system appear stronger than it really is. For LE data, this is especially problematic because deployment almost always involves some form of distribution shift. The evaluation philosophy of this dissertation is therefore intentionally conservative: whenever possible, models should be assessed under conditions that reflect temporal, cohort, contextual, or institutional shift rather than only under i.i.d. splits.

This stance influences all empirical chapters. In the early educational forecasting work, limited-data conditions and early-in-semester prediction windows were treated as first-class constraints rather than as incidental details [26,27]. In the later cross-distribution chapters, benchmark design centers on separating ID from OOD testing so that the question becomes whether the model has learned a transfer-

able behavioral representation rather than a source-specific mapping [3, 16, 21, 22]. This is why the dissertation repeatedly distinguishes between performance on a held-out subset of the training period and performance on a future or otherwise shifted period. The former indicates whether the model can fit within the source environment. The latter is closer to the actual scientific question.

A second principle concerns the relation between *prediction* and *interpretation*. In human-centered settings, a model may be practically more useful if its output carries interpretable reasoning structure rather than merely a score. This is one reason the dissertation increasingly favors narrative generation and semantically meaningful intermediate representations. The point is not that narratives automatically make systems transparent in a strong causal sense. Rather, narratives can force the modeling pipeline to operate at a level of abstraction closer to how human analysts think about trajectories: not merely as vectors, but as evolving stories with salient transitions, contextual modifiers, and likely future paths. The narrative output space can therefore act as a regularizing bottleneck and as an interpretive interface. This is especially important in Chapter 3 and beyond, where generated prospective narratives are used both as outputs and as evidence that the model has learned a semantically coherent view of the trajectory [3].

A third principle is that missingness should be evaluated not only by imputation fidelity but also by downstream utility and robustness. In many traditional settings, imputation quality is measured by reconstruction error with respect to held-out observed values. That metric is useful, but it does not fully capture the role of missingness in LE data. An imputed value can be numerically plausible yet semantically unhelpful or even misleading. The dissertation therefore treats downstream forecasting performance under varying missingness conditions as an essential part of evaluation. CRILM and later chapters emphasize that the value

of an imputation strategy lies in how well it preserves or recovers behaviorally meaningful context for the downstream task [1].

A fourth principle is that reliability must be measured explicitly. For a probabilistic model or a classifier with confidence outputs, it is not enough to know average accuracy. It is also important to know whether high-confidence predictions are actually more likely to be correct, whether the model becomes overconfident under shift, and whether uncertainty signals deteriorate gracefully as stress increases. Reliability diagrams, expected calibration error, confidence histograms, and OOD sensitivity are therefore conceptually central to the dissertation, particularly in the PRISM chapter [23–25]. Calibration is not treated as an optional appendix metric; it is part of what it means for a system to be usable in intervention settings.

A fifth principle is that representation quality should be assessed indirectly through ablation and comparison, not only through intuition. Because the dissertation argues that representation design is central, it becomes necessary to show that different representations of the same underlying trajectory yield materially different generalization behavior. This is why the empirical chapters compare raw serialization, contextualized verbalization, meta-narratives, unimodal visual encodings, multimodal combinations, and frozen-prior variants. The goal is not simply to find the best number on a benchmark, but to make the case that certain representational choices preserve more transferable information than others. LE-Viz, for example, is important not only because it improves performance, but because its ablations show that spatial organization itself contributes to the gain [21]. PRISM is important not only because it achieves strong OOD results, but because it links those gains to complementary streams and fixed coherence constraints [22].

Finally, the evaluation philosophy of the dissertation is inseparable from its view of foundation models. A large pretrained model is not treated as an oracle

whose size guarantees success. It is treated as a source of prior structure that can either help or hurt depending on how the task is framed, how the data is represented, and how adaptation is controlled. This leads to a methodological position that is neither purely task-specific nor purely scaling-centric. The dissertation repeatedly shows that modestly sized or efficiently adapted models can outperform larger but poorly aligned baselines when the representational interface is well designed. The lesson is not that scale is irrelevant, but that scale without alignment to the structure of LE data is insufficient.

These principles together explain why the dissertation places so much emphasis on chapter-level bridges, ablations, and progressively stronger evaluation regimes. The scientific claim is cumulative: if each chapter solves a problem exposed by the previous one, then the evidence for the dissertation lies not only in the final architecture but in the path by which each methodological choice becomes necessary. The introduction therefore frames the dissertation not simply as a set of empirical wins, but as a structured argument about how one should think about LE modeling in the first place.

1.14 Summary of Dissertation Contributions

The dissertation's contributions can be understood as stages in a single evolving argument.

1.14.1 Contribution 1: Reframing early LE forecasting through contextualized language models

The earliest work in the dissertation shows that early forecasting of academic performance can be substantially improved by formulating the task for pretrained

language models and by enriching the textual representation with contextual information [26, 27]. Rather than relying only on early cognitive scores, the representation incorporates distal and proximal factors so that the model interprets student trajectories in context. This establishes the dissertation's initial claim that representation quality can matter more than task-specific numeric architecture when data is scarce.

1.14.2 Contribution 2: Broadening LE modeling beyond purely cognitive trajectories

Subsequent work expands from narrowly cognitive educational forecasting to broader LE data in education, incorporating qualitative and non-cognitive dimensions that are central to understanding student development [20]. This contribution is important because it defines LE data not as an edge case but as a legitimate and underexplored machine learning object whose richness exceeds what traditional educational prediction pipelines capture.

1.14.3 Contribution 3: Modeling missingness as part of the signal

CRILM introduces a context-aware language-model-based approach to imputation, demonstrating that semantically meaningful descriptor generation can improve downstream performance and robustness in the presence of missing data [1]. This contribution reorients missing-data handling from numeric patching toward representational enrichment and is especially relevant to LE data, where missingness often reflects meaningful behavioral processes.

1.14.4 Contribution 4: Building LLM frameworks for qualitative longitudinal forecasting

The dissertation develops full NLP-based pipelines for forecasting engagement from qualitative LE data, including LLM-informed imputation, feature selection, and fine-tuning [2]. This shows that the approach is not limited to mixed cognitive datasets; it can also handle predominantly qualitative and sparse longitudinal responses where traditional numeric models perform poorly.

1.14.5 Contribution 5: Advancing cross-distribution generalization through narrative-based representations

ConText-LE contributes the dissertation's first explicit OOD-focused framework for LE forecasting [3]. Its Meta-Narrative representation and Prospective Narrative Generation formulation demonstrate that the way trajectories are summarized and the way outputs are formulated both matter for transfer. This chapter turns the dissertation clearly toward cross-distribution generalization as the central evaluation criterion.

1.14.6 Contribution 6: Introducing structure-preserving multimodal LE modeling

LE-Viz extends the representational argument beyond text by showing that visual encodings of feature–time structure can complement narrative representations and improve cross-distribution performance [21]. The significance of this contribution lies not merely in adding another modality, but in identifying the specific structural information that text serialization loses and visual encoding can preserve.

1.14.7 Contribution 7: Constraining learning with frozen priors for reliable behavioral generalization

PRISM unifies the dissertation's ideas into a frozen-backbone multimodal framework that emphasizes complementary views, multi-resolution supervision, and a frozen semantic coherence constraint [22]. This contribution is the dissertation's final methodological step because it links representation quality to reliability and argues that transferable learning requires explicit constraints on adaptation.

Collectively, these contributions form a coherent progression: from contextualized language interfaces, to semantically aware treatment of missingness, to narrative-based OOD generalization, to structure-preserving multimodality, and finally to frozen-prior constraint-based learning for reliable transfer.

1.15 Dissertation Organization

The dissertation is organized by intellectual dependency rather than by publication chronology. The chapters are arranged to show how each methodological step arises from a limitation exposed by the previous one.

Chapter 2: Traditional ML and DL Approaches for LE data

Chapter 2 establishes the baseline problem setting. It reviews traditional machine learning and deep learning approaches used for LE forecasting, especially in educational and behavioral applications, and shows why these methods struggle when the data is small, heterogeneous, and shift-prone. The chapter clarifies the assumptions that motivate the later transition to language-centric approaches.

Chapter 3: LLMs for Contextual and Narrative Modeling of LE Data

Chapter 3 introduces the language-modeling stage of the dissertation. It begins with early work on personalization and contextualization for academic forecasting and then expands to broader LE trajectories. It develops the idea that verbalization and contextual enrichment provide a more semantically aligned interface for pretrained models. It also introduces the move from direct classification to narrative forecasting, culminating in ConText-LE.

Chapter 4: Vision-Language Modeling and Multimodal Structure Preservation

Chapter 4 argues that language-centered approaches, although powerful, still suffer from structural loss due to serialization. It introduces LE-Viz and develops the multimodal argument that text and visual encodings preserve complementary invariants of the same trajectory. The chapter demonstrates how structure-preserving visual representations can improve cross-distribution generalization beyond text-only models.

Chapter 5: PRISM and Frozen-Prior Multimodal Constraints for Reliability

Chapter 5 presents PRISM as the dissertation's culminating framework. It shows how complementary temporal, spectral, and semantic streams can be fused under frozen priors and multi-objective constraints to improve both accuracy and reliability under distribution shift. The chapter positions frozen semantic coherence as a mechanism for discouraging shortcut learning and promoting transferable behavioral reasoning.

Chapter 6: Conclusion and Future Directions

The final chapter synthesizes the dissertation's findings into a set of design principles for future LE modeling systems. It reflects on implications for human-centered AI, reliable foundation models, domain-shift evaluation, and multimodal learning under scarce-data regimes. It also outlines open research questions, including uncertainty-aware intervention design, causal interpretation of behavioral trajectories, and broader scientific applications of foundation models to longitudinal human data.

1.16 Bridge to the Remainder of the Dissertation

This introduction has argued that LE modeling is best understood as a generalization-first problem defined by contextual dependence, heterogeneity, missingness, structural complexity, and deployment under shift. It has further argued that each methodological transition in the dissertation follows from a concrete representational limitation: classical ML/DL pipelines flatten meaning; text-based LLM approaches recover semantics but collapse structure; multimodal models recover structure but require stronger constraints to remain reliable; frozen priors help provide those constraints.

The chapters that follow develop this argument in detail. They show that the central scientific challenge is not how to apply the largest available model to human-centered longitudinal data, but how to design representations and learning procedures that respect what this data actually is: temporally evolving, semantically rich, context-bound, partially observed, and deployed in environments that differ from the one in which the model was trained. The dissertation contribution is therefore not a single architecture, but a connected research program whose end

result is a new way of thinking about LE data itself—as an object that demands multimodal, semantically grounded, and reliability-aware AI.

Chapter 2

Traditional Machine Learning and Deep Learning Approaches for LE data

2.1 Introduction

Before the dissertation turns to language models, multimodal learning, and frozen-prior architectures, it is necessary to establish a rigorous baseline understanding of what traditional ML and DL methods can and cannot do for LE data. This chapter develops that foundation.

At first glance, LE forecasting appears to be a familiar supervised learning problem. Given a window of prior observations about an individual, one may attempt to predict a later state, such as academic performance, engagement, affective change, depression risk, or well-being. This framing naturally invites the use of standard tabular learners, sequence models, and multivariate time-series architectures. In education, models built from attendance records, assessment histories, and behavioral traces have long been used for early warning and performance prediction [8–10]. In behavioral sensing and digital health, similar strategies have been used to forecast stress, depression, and related outcomes from passively sensed signals and self-reports [7, 16]. From this perspective, it is entirely reasonable that the first stage of this dissertation began from traditional ML and DL methods.

Yet the main argument of this chapter is that LE data is not simply another instance of multivariate time-series classification. The difficulty is not only temporal dependence. It is the combination of temporal dependence with context-sensitive meaning, heterogeneous modalities, qualitative self-reports, missing-not-at-random patterns, small-sample regimes, and cross-distribution deployment. In such settings, the central weakness of traditional ML/DL approaches is not simply insufficient model size. It is a deeper mismatch between their representational assumptions and the informational structure of LE data.

This mismatch becomes visible throughout the empirical arc of the dissertation. In early educational forecasting, conventional numeric baselines such as LSTM, CNN, Transformer, and SVM remain intuitive first choices, but they underperform once contextual student factors become important [26, 27]. In richer longitudinal educational datasets containing qualitative non-cognitive trajectories, background factors, and substantial missingness, the gap widens further because numeric representations fail to preserve the behavioral meaning of the sequence [20]. In sparse qualitative engagement forecasting, traditional baselines become even more unstable, often showing positive-class bias and weak macro-F1 despite access to the same underlying trajectory content [2]. Finally, in cross-distribution behavioral forecasting, the GLOBEM benchmark made the problem unmistakable by showing that a broad range of traditional and domain-generalization methods fail to exceed 52.8% OOD accuracy [3, 16, 17, 22].

This chapter therefore serves four purposes. First, it formalizes the traditional predictive view of LE data and clarifies the assumptions embedded in that view. Second, it reviews the major families of classical ML and DL methods that have been used for LE forecasting, including tabular models, recurrent networks, convolutional sequence models, modern time-series transformers, and

missing-data-aware temporal models. Third, it uses the dissertation's own empirical case studies to show where these approaches remain useful and where they fail. Fourth, it establishes the representational and generalization gaps that motivate the transition to LLMs in Chapter 3.

2.2 LE data Under a Traditional Predictive Lens

Traditional predictive pipelines usually begin by assuming that each individual can be represented either as a fixed feature vector or as a sequence of predominantly numeric feature vectors. Under this view, the trajectory of an individual i across T time steps is written as

$$X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,T}\}, \tag{2.1}$$

where $x_{i,t} \in \mathbb{R}^F$ is the observed feature vector at time step t. A forecasting model then learns

$$\hat{y}_{i,t+h} = f_\theta(x_{i,t-w+1}, \ldots, x_{i,t}), \tag{2.2}$$

for some lookback window w, horizon h, and parameters θ.

As a mathematical abstraction, this formulation is unobjectionable. The challenge lies in the implicit assumptions that typically accompany it. In practice, traditional ML/DL pipelines for LE data often assume that:

1. each feature has a stable, context-independent meaning;
2. the relevant predictive structure can be expressed numerically;
3. missing values are noise to be repaired rather than behavior to be interpreted;
4. flattening, aggregation, or serialization are neutral preprocessing steps;
5. strong ID performance is a reasonable proxy for real-world usefulness.

These assumptions are often acceptable for clean industrial telemetry or stationary physical signals. They are much less reliable for LE data. In educational and behavioral settings, variables such as confidence, engagement, sleep, mobility, self-evaluation, or identity do not behave like independent physical sensors. Their meaning often depends on timing, social context, background factors, and surrounding observations. A decline in one variable may be a risk signal in one context and an adaptive response in another. This is why LE modeling is not simply about learning from temporal order. It is about learning from *temporal order plus situated meaning*.

A second issue is representational compression. A standard preprocessing step is to flatten a feature–time matrix into a vector:

$$X_i \in \mathbb{R}^{T \times F} \quad \longrightarrow \quad \mathrm{vec}(X_i) \in \mathbb{R}^{TF}, \tag{2.3}$$

or to aggregate each feature into summary statistics:

$$z_i = \psi(X_i), \qquad \hat{y}_i = h(z_i), \tag{2.4}$$

where $\psi(\cdot)$ may compute means, slopes, counts, minima, maxima, or handcrafted trend descriptors. These transformations are convenient, but they are not information-neutral. They often erase temporal locality, obscure cross-feature synchrony, and merge trajectories that are behaviorally distinct yet numerically similar in aggregate.

The traditional predictive lens is therefore both necessary and incomplete. It supplies the baseline formalism, but it also introduces the very assumptions that the rest of the dissertation will systematically challenge.

2.3 Why Traditional Methods Were the Natural Starting Point

The early stages of this dissertation entered LE modeling through educational forecasting. This was a natural point of entry for several reasons. Educational outcomes offer clearly defined targets, the intervention motivation is immediate, and the available data already contains a useful mixture of cognitive, non-cognitive, and background information. Moreover, the educational setting makes the temporal challenge concrete: if one wishes to identify students who are struggling, useful predictions must be made early, when the available trajectory is still short and incomplete [26, 27].

Under those conditions, traditional methods are the obvious first baseline. If early-semester assessment histories can forecast end-of-semester outcomes, then recurrent networks, convolutional sequence models, support vector machines, and shallow transformers are reasonable tools to try. They are efficient, well understood, and supported by a large body of literature. Even when they fail, they fail informatively: their error modes reveal which aspects of the trajectory cannot be captured by a purely numeric or context-agnostic formulation.

That is precisely what happened in the empirical trajectory of this dissertation. In early academic performance forecasting, numeric baselines remained important comparators, but the introduction of contextual student information exposed their limitations. In subsequent work on richer educational LE datasets, the challenge shifted from numeric trend extraction to behavioral interpretation under qualitative, sparse, and heterogeneous conditions [20]. These transitions were not arbitrary methodological choices. They were empirical consequences of how the data itself resisted the assumptions of standard ML/DL pipelines.

2.4 Classical Machine Learning Foundations for LE Forecasting

2.4.1 Engineered features and shallow supervised learning

The oldest and still one of the most common approaches to longitudinal behavioral forecasting is to engineer features from the trajectory and then apply a shallow supervised learner such as logistic regression, support vector machines, decision trees, random forests, or boosting ensembles. This design remains attractive because it works well with modest sample sizes, permits explicit feature construction, and often produces models that are easier to inspect than deep networks.

In educational settings, such features may include averages of assessment scores, counts of missed assignments, attendance rates, slope estimates over time, or summary statistics derived from learning-management-system activity [8–10]. In behavioral sensing, similar summaries can be extracted from sleep, mobility, communication patterns, and device usage [7, 16]. The underlying assumption is that a sufficiently rich handcrafted vector can compress the trajectory without sacrificing predictive content.

This assumption has three practical strengths. First, it reduces the sample complexity of the learning problem by replacing long sequences with lower-dimensional summaries. Second, it allows the modeler to encode prior beliefs about which temporal properties matter. Third, it creates a computationally lightweight pipeline that is suitable for small datasets and rapid deployment. These advantages explain why shallow models continue to appear in LE applications even today.

2.4.2 Support vector machines and margin-based classification

Among classical baselines, support vector machines (SVMs) are especially common when the goal is categorical forecasting from small or medium-sized datasets. SVMs seek a separating hyperplane with maximal margin:

$$f(x) = \text{sign}(w^\top x + b), \tag{2.5}$$

with optimization over w and b regularized by the margin constraint. In practice, SVMs can perform surprisingly well when the feature space is informative and the dataset is not too large. This is why SVMs were included among the early educational baselines in this dissertation.

Indeed, in the earliest academic forecasting experiments, the SVM remained competitive on the shortest forecasting window. Using only cognitive features, it reached 59% accuracy in the 2-week setting and remained one of the stronger numeric baselines alongside the numeric Transformer [26]. This result is important because it shows that traditional methods are not uniformly weak. When the signal is relatively structured and the task is short-horizon classification, margin-based classifiers can still extract useful predictive patterns.

At the same time, the SVM result also reveals a ceiling. Its performance plateaued as the window length increased, whereas richer contextualized representations supported stronger gains by models that could reason over broader information [26]. In other words, the SVM was able to exploit readily separable structure, but not the deeper contextual dependencies that emerged when the task moved beyond cognitive-only numeric trajectories.

2.4.3 Tree ensembles and tabular robustness

Tree-based methods such as Random Forest [35] are another natural baseline in LE tasks because they can model non-linear relationships, tolerate modest noise, and often remain robust when feature scales differ. In sparse or heterogeneous tabular settings, they are frequently stronger than one might expect from their conceptual simplicity.

This pattern reappears later in the dissertation. In qualitative engagement forecasting, Random Forest becomes the strongest numeric baseline for one of the engagement dimensions (Performance Self-Evaluation), reaching 53.5% balanced accuracy and 52.5% macro-F1, even as other numeric models remain near chance [2]. That result matters because it confirms that the shortcomings of traditional methods do not stem from a lack of care in baseline selection. Strong tabular learners were tried. They still struggled once the forecasting problem depended on qualitative interpretation, contextual nuance, and behaviorally meaningful missingness.

2.4.4 The core limitation of feature engineering

The deepest problem with classical feature engineering is not that it always performs poorly. It is that it assumes the trajectory can be hand-compressed into a stable set of summary statistics. Formally, the aggregation operator $\psi(\cdot)$ in Equation 2.4 is usually many-to-one: multiple distinct trajectories can map to the same or nearly the same summary vector. This is acceptable only when the discarded differences are irrelevant to the task. In LE data, they often are not.

Two students may have similar average confidence over four weeks, yet one may be gradually improving while the other is oscillating sharply. Two participants

may share the same mean sleep duration, yet one may have a stable rhythm while the other alternates between severe restriction and recovery. Two students may show the same rate of missing responses, but in one case the missingness may reflect disengagement while in the other it reflects scheduling noise. Once such distinctions matter, engineered summaries begin to underfit the experiential structure of the data.

2.5 Deep Learning Approaches

2.5.1 Recurrent sequence models

The first major deep learning response to the limitations of aggregate tabular models was to retain sequence order explicitly. Recurrent neural networks (RNNs) and their gated variants, especially long short-term memory networks (LSTMs) and gated recurrent units (GRUs), were designed for exactly this purpose [33, 44, 45]: to update a hidden state as new observations arrive over time. A generic recurrent update can be written as

$$h_t = \phi(W_h h_{t-1} + W_x x_t + b), \tag{2.6}$$

where h_t is the hidden state at time t, x_t is the input, and ϕ is a nonlinear activation. Gated variants such as the LSTM improve stability by controlling what information is retained, forgotten, and written to memory.

For LE data, recurrent models are appealing because they preserve temporal order without requiring the analyst to specify all temporal features manually. They can, in principle, detect whether a trajectory is rising, falling, oscillating, or transitioning abruptly. This is why LSTMs were used repeatedly as baseline

models in the educational forecasting stages of the dissertation [2, 26].

However, recurrent models inherit the assumptions of their input representation. If the input is a sequence of numeric feature vectors, then the model can learn temporal correlations among those vectors, but it still does not know what the values *mean* in context. It can observe that confidence dropped, but not whether that drop should be interpreted as disengagement, healthy recalibration, or performance-related realism. Moreover, recurrent models are often fragile in small-data settings and are sensitive to irregular sampling and missing values.

The dissertation's results expose these weaknesses directly. In early academic forecasting, the LSTM remained well below the strongest contextualized language-based methods [26]. In qualitative engagement forecasting, the LSTM was the best numeric baseline in several dimensions, but even then it peaked at only 55.5% balanced accuracy and 54.0% macro-F1, with a much lower overall mean than textual models [2]. More importantly, the numeric baselines frequently showed class imbalance pathologies, particularly a tendency to predict the positive class excessively. Thus, preserving order is not enough when the primary challenge is semantic interpretation of qualitative and sparse trajectories.

2.5.2 Convolutional sequence models

One-dimensional convolutional neural networks (1D CNNs) offer another traditional way to model temporally ordered data [46,47]. Rather than maintaining a recurrent state, they apply learned filters over local neighborhoods in time. This makes them efficient and often effective when local motifs matter more than long-range dependencies.

In LE forecasting, 1D CNNs can detect short-term temporal patterns such as bursts, dips, or local trend changes. They also train faster than many recurrent

architectures. These strengths made them useful baselines in the dissertation's early experiments [2, 26]. Yet their limitations are closely related to those of other numeric models. Convolution can only exploit patterns that are already represented in an aligned numeric grid. It does not solve the problem of contextual meaning, and it does not naturally reconcile heterogeneous modalities whose signals become predictive only when jointly interpreted.

The engagement forecasting experiments provide a clear example. Although the 1D CNN was occasionally competitive on one dimension when all non-cognitive features were included, it remained substantially weaker than RoBERTa-style textual models and did not overcome the broader limitations of numeric input conversion [2]. This is a recurring theme in the chapter: traditional models may capture one piece of the problem while still missing the aspect that matters most.

2.5.3 Transformer-based time-series models

The next major step in traditional DL for temporal data was the adoption of attention-based architectures. Transformers replace recurrent updates with pairwise interaction through self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.7}$$

where Q, K, and V denote learned query, key, and value projections. For multivariate time-series, this allows the model to capture long-range dependencies and variable interactions more flexibly than many recurrent alternatives.

Transformer-derived architectures such as PatchTST and iTransformer have become strong baselines in general time-series forecasting because they handle long

input windows, exploit global dependencies, and often scale better than classic recurrent models [48, 49]. Within the dissertation, both families were considered in the later LE-Viz and PRISM comparisons to ensure that modern numeric sequence modeling, not only older recurrent baselines, was adequately represented [21, 22].

Their inclusion is important because it rules out an overly simple explanation for baseline failure. The problem is not merely that the wrong deep architecture was chosen. Even with modern attention-based time-series models, OOD performance on GLOBEM remained weak: approximately 49.88% OOD accuracy for PatchTST and 51.06% for iTransformer in the LE-Viz comparisons [21]. This level of performance is only marginally above chance and far below the stronger multimodal approaches developed in later chapters. The implication is clear: better sequence architecture alone does not fix the representational mismatch between traditional numeric modeling and LE data.

2.5.4 Missing-data-aware temporal models

Because LE datasets are often incomplete, another strand of traditional DL work focuses on explicitly modeling missingness. Two representative examples are GRU-D and BRITS [31, 32]. These models augment recurrent processing with missingness masks, time gaps, and learned imputation dynamics, thereby improving over naive mean imputation or forward filling when values are missing irregularly.

Conceptually, these methods are highly relevant to LE data because they acknowledge that incompleteness is part of the temporal process. Yet even here, a limit appears. Missing-data models like GRU-D and BRITS treat missingness primarily as a numerical state to be repaired or compensated for. They do not, by themselves, attach *semantic meaning* to absence. In LE settings, especially with

self-reported non-cognitive measures, a missing response can carry contextual information about disengagement, burden, avoidance, or uncertainty. This is one reason why later chapters of the dissertation move beyond purely numerical imputation toward context-aware language-based strategies [1, 2].

For the purposes of this chapter, the important point is not that GRU-D or BRITS are ineffective. Rather, they show the limit of what traditional temporal DL can achieve while remaining inside a fundamentally numeric representational regime. They improve handling of incomplete sequences, but they still do not transform absence into interpretable behavioral signal.

2.5.5 Domain generalization methods

One might reasonably ask whether the limitations above can be solved not by changing the representation, but by adopting stronger training strategies designed for robustness under distribution shift. Domain adaptation and domain generalization offer exactly that promise [17, 18, 50, 51]. Methods in this family aim to learn representations that transfer across environments by aligning domains, enforcing invariance, or discouraging reliance on spurious features.

This literature is highly relevant to the dissertation because the central deployment challenge of LE data is cross-cohort and cross-temporal generalization. Yet the GLOBEM benchmark showed that, in this domain, the problem is not easily solved by applying domain-generalization methods off the shelf. Nine published depression-detection algorithms and eight domain generalization approaches all failed to exceed 52.8% OOD accuracy [16, 22]. This result is foundational to the dissertation's argument. It demonstrates that even when robustness is made the explicit optimization target, traditional representations still fail to capture the contextual and experiential structure needed for transfer.

Table 2.1: Summary of traditional method families for LE forecasting. The table clarifies why these methods remain important baselines while also indicating the specific points at which they begin to fail for contextual, qualitative, and cross-distribution LE tasks.

Method family	**Typical input form**	**What it captures well**	**Principal blind spot in LE data**
Linear / margin-based models	Engineered tabular summaries	Small-data robustness, simple decision boundaries, interpretable feature effects	Depend heavily on handcrafted summaries; weak at preserving temporal and contextual nuance
Tree ensembles	Engineered tabular summaries	Non-linear tabular interactions, moderate robustness to noisy variables	Still collapse trajectories into feature vectors and do not attach meaning to time or missingness
RNN / LSTM / GRU	Ordered numeric sequences	Sequential dependence, short- to medium-range temporal patterns	Context-sensitive semantics remain latent; fragile under small data, irregularity, and qualitative inputs
1D CNN / hybrid temporal models	Aligned numeric sequences	Local motifs and efficient pattern extraction	Capture local signal but not deeper behavioral interpretation or heterogeneous modality alignment
Time-series transformers	Long multivariate numeric sequences	Global dependency modeling, flexible variable interactions	Strong sequence learners, but still bounded by the limitations of homogeneous numeric representation
Missing-data-aware temporal models	Numeric sequences with masks and time gaps	Better numerical handling of incompleteness	Treat missingness mainly as a repair problem rather than as semantically meaningful behavior
Domain-generalization methods	Environment-labeled numeric representations	Explicit robustness objective under distribution shift	Invariance objectives help, but cannot fully compensate for impoverished representations

2.6 Empirical Stress Tests from the Dissertation

This section uses three case studies from the dissertation to show how the limitations of traditional ML/DL methods become visible as the LE problem becomes progressively richer.

2.6.1 Case Study I: Early academic performance forecasting

The first stress test comes from early academic performance forecasting. In this setting, students' end-of-semester performance is predicted using only the first 2, 4, or 8 weeks of data [26]. Traditional baselines included an SVM, LSTM, CNN, and numeric Transformer trained on cognitive features. This task is important because it is, in many ways, a relatively favorable environment for traditional methods: the data is structured, the target is clearly defined, and the cognitive component is naturally numeric.

Even so, the baseline results were mixed. The SVM achieved 59% accuracy in the 2-week forecasting setting, and the numeric Transformer reached 55%, both slightly ahead of the LLM when only cognitive features were considered [26]. However, that advantage did not scale with richer context or longer windows. Once the trajectory was contextualized with proximal non-cognitive and background information, performance improved substantially for the richer representations, with a fully contextualized model reaching 77% accuracy at week 2 and 89% accuracy by week 8 [27]. The lesson is not that numeric baselines are useless. It is that they saturate early because they do not reinterpret the trajectory through context.

Figure 2.1 is helpful because it shows the precise point at which traditional methods stop being enough. They are not invalid baselines. They are simply

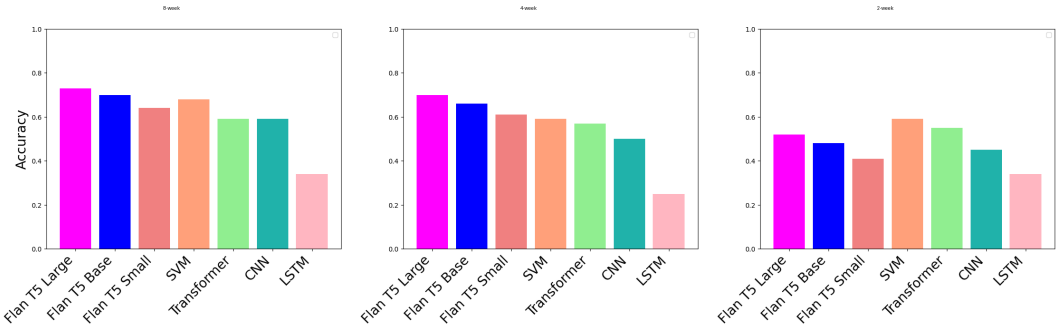


Figure 2.1: Comparison between traditional numeric baselines and language-centered models in early academic performance forecasting across 8-week, 4-week, and 2-week settings. The plots are adapted from the user's earlier forecasting study and are included here because they make visible a recurring pattern in this dissertation: traditional baselines can remain competitive in the shortest and most structured settings, but they struggle to scale as the forecasting problem becomes more contextual and semantically rich [26].

insufficient once the forecasting problem requires more than numeric pattern recognition.

2.6.2 Case Study II: Rich educational LE trajectories

The second stress test comes from the dissertation's richer educational LE dataset, which represents a holistic view of students' academic trajectories through three groups of factors: 41 quantitative cognitive dimensions, 28 repeated qualitative non-cognitive dimensions, and 9 background dimensions [20]. This dataset is especially useful for Chapter 2 because it makes the representational problem concrete.

From the perspective of traditional ML/DL, the dataset introduces several compounding difficulties. The number of students is modest ($N = 48$), the time-varying dimensions are heterogeneous, the non-cognitive signals are qualitative, and the missing-value rate is substantial [20]. These characteristics break the tacit assumptions that underlie many standard sequence models. The challenge is no

longer just to model a short numeric trajectory. It is to relate multiple partially observed views of the student over time, some numeric, some qualitative, and some static.

This dataset marks an important conceptual turning point in the dissertation. At this point, the problem can no longer be plausibly described as ordinary multivariate forecasting. It becomes clear that the dominant bottleneck is not architecture selection but representation design. Traditional numeric models can ingest more variables, but they still treat those variables as values rather than as contextually meaningful descriptors.

2.6.3 Case Study III: Qualitative engagement forecasting

The third stress test comes from forecasting student engagement dimensions from qualitative longitudinal data [2]. Here, the input is not dominated by cognitive scores but by sparse non-cognitive self-reports collected over time. Four engagement dimensions are forecasted, and the numeric baselines include Random Forest, SVM, 1D CNN, LSTM, and Transformer.

This setting is especially revealing because the trajectory content itself is closer to language than to continuous sensor measurements. To fit traditional models, qualitative responses must be converted into scores, missing values must be numerically repaired, and the resulting sequence must be presented as a fixed tensor. In that pipeline, multiple layers of behavioral meaning are compressed before learning even begins.

The results show the cost of that compression. Across the four engagement dimensions, the numeric baselines average only 50.8% balanced accuracy and 46.9% macro-F1 [2]. The strongest traditional model, LSTM, leads only in selected dimensions and still remains far below the best textual encoders. The baseline

models also show a tendency toward positive-class bias, making them unreliable in practice even when headline accuracy does not look catastrophic. This is a critical finding for the dissertation because it shows that, for sparse qualitative LE data, traditional ML/DL approaches are not just suboptimal; they are often behaviorally misaligned.

2.6.4 Case Study IV: Cross-distribution behavioral forecasting

The final stress test comes from the cross-distribution setting. The GLOBEM benchmark was created precisely to evaluate whether behavioral models can transfer across time periods and institutions [16]. This is the real test of whether an LE model is learning robust structure or source-specific shortcuts.

The benchmark findings are stark. A wide range of published behavioral modeling methods, together with multiple domain-generalization baselines, failed to exceed 52.8% OOD accuracy [3, 16, 22]. Later in the dissertation, more advanced approaches raise this number substantially, but that only reinforces the present point: traditional ML/DL methods, even when paired with robustness-oriented training strategies, do not naturally discover the distribution-invariant structure required for transfer in LE data.

Table 2.2: Representative empirical stress tests showing where traditional ML/DL methods begin to break under increasingly realistic LE conditions. The purpose of the table is not to claim that all traditional methods always fail, but to show the specific conditions under which their assumptions become misaligned with the task.

Setting	**Traditional baselines examined**	**Representative strongest numeric result**	**What the result reveals**
Early academic performance forecasting [26, 27]	SVM, CNN, LSTM, numeric Transformer	SVM reaches 59% accuracy at 2 weeks; numeric Transformer 55%	Traditional methods can exploit short, structured cognitive signals, but performance saturates and does not benefit from contextual student information as strongly as richer representations.
Rich educational LE trajectories [20]	Conventional numeric modeling becomes difficult because the data mixes 41 cognitive, 28 qualitative non-cognitive, and 9 background dimensions	No traditional numeric formulation cleanly preserves all factors	The problem shifts from sequence fitting to representation design; small sample size, qualitative signals, and missingness strain standard ML/DL assumptions.
Qualitative engagement forecasting [2]	Random Forest, SVM, 1D CNN, LSTM, Transformer	Best traditional results peak around 55.5% balanced accuracy and 54.0% macro-F1; average baseline performance is lower	Once the trajectory is sparse, qualitative, and missing-not-at-random, numeric conversion introduces severe information loss and unstable class behavior.
GLOBEM cross-distribution forecasting [16, 22]	Published behavioral models and domain-generalization methods	Best reported OOD accuracy remains 52.8%	Even explicit robustness training does not overcome representational mismatch; the core issue is not only learning objective but also how LE trajectories are represented.

2.7 Where Traditional ML and DL Methods Fail

The empirical case studies above suggest that the weaknesses of traditional methods are not accidental. They arise from structural properties of LE data. This section organizes those weaknesses into five major failure modes.

2.7.1 Failure 1: context-independent semantics

Perhaps the most fundamental problem is that traditional methods usually treat feature values as having stable meaning across individuals and contexts. This assumption is often harmless in physical systems, but it is deeply problematic in human-centered forecasting. The significance of a sleep change, confidence drop, or engagement dip depends on surrounding context, personal baseline, and stage in the semester or life event. Traditional ML/DL systems can learn correlations between variables and outcomes, but they do not inherently interpret those variables relative to lived context.

This issue parallels a historical limitation in NLP. Early word embeddings represented words as static vectors regardless of context, until contextualized encoders such as ELMo and BERT demonstrated that meaning is inherently conditional on surrounding text [52, 53]. A similar shift is needed for LE data. A variable such as low motivation is not a fixed token with a universal meaning. Its interpretation depends on the trajectory around it.

2.7.2 Failure 2: representational collapse

Traditional pipelines often flatten or summarize LE trajectories too aggressively. Classical tabular models collapse time into engineered summaries. Recurrent and convolutional models preserve order, but still operate on a homogeneous numeric

stream. Time-series transformers retain more global dependency structure, yet they still assume that the sequence has already been encoded in an adequate numeric form.

None of these methods naturally preserve all three of the following at once:

1. semantic interpretation of qualitative content,
2. cross-feature relational structure,
3. temporal behavioral dynamics.

Once LE data contains qualitative self-reports, background context, and meaningful absence, this representational collapse becomes a central source of failure. The later multimodal chapters of the dissertation will argue that no single narrow representation suffices and that complementary modalities are needed to preserve the different invariants present in the trajectory [21, 22]. For now, the critical point is that the collapse begins long before multimodality enters the picture. It is already present in standard numeric ML/DL pipelines.

2.7.3 Failure 3: missingness is treated as nuisance rather than signal

Traditional methods are usually designed around the idea that missing values should be imputed, masked, or ignored. This is reasonable when missingness is light and approximately random. LE data often violates that assumption. In repeated self-report settings, skipped responses can correlate with disengagement, emotional burden, or avoidance. In sensing settings, missingness can reflect device habits, wear patterns, or behavioral disruption.

Numeric imputation methods improve continuity, but they generally do not enrich missingness with behavioral meaning. As a result, the repaired sequence

often looks mathematically cleaner while becoming semantically poorer. This gap eventually motivates the dissertation's move to context-aware imputation and narrative modeling [1, 2], but the motivation is already visible here: traditional ML/DL treats absence as a defect in the data matrix, not as part of the experiential state.

2.7.4 Failure 4: small-data instability and shortcut learning

LE datasets are frequently expensive to collect, which means that model complexity and data availability are often badly mismatched. When N is small relative to the dimensionality and heterogeneity of the trajectory, traditional deep architectures are vulnerable to overfitting. Even when explicit overfitting is controlled, the model may learn shortcuts that work inside the source distribution without reflecting stable behavioral structure.

This problem is especially acute when the dataset combines static, dynamic, cognitive, non-cognitive, and qualitative variables. The richer the input, the easier it is for the learner to pick up unstable associations. In such settings, simply adding architectural sophistication does not ensure better generalization. It may instead enlarge the space of spurious solutions.

2.7.5 Failure 5: poor cross-distribution generalization

The final and most consequential failure mode is weak transfer under distribution shift. A traditional model may achieve reasonable held-out accuracy within a single cohort and still fail dramatically when evaluated on later years, new institutions, or new participant groups. The GLOBEM results are the clearest demonstration of this problem [16]. But similar patterns also motivate the educational forecasting work in this dissertation, where the goal is not only to fit one

course offering but to build methods that remain useful as cohorts, contexts, and behavior patterns change [3, 22].

Cross-distribution failure is important because it reveals what the model has actually learned. A model that transfers poorly has likely captured source-specific associations rather than deeper causal or semantically grounded structure. This diagnosis is central to the dissertation. It is one reason the later chapters emphasize narrative reasoning, multimodal complementarity, and frozen-prior constraints rather than only larger neural architectures.

2.8 What Traditional Methods Still Contribute

Although this chapter has been critical of traditional ML and DL methods, it would be a mistake to dismiss them. They remain important for at least four reasons.

First, they are scientifically necessary baselines. Without strong numeric baselines, it would be impossible to show that later language-based and multimodal methods are solving a real problem rather than benefiting from weak comparisons. Throughout this dissertation, traditional models provide precisely that baseline discipline.

Second, they are often operationally attractive. In settings where the trajectory is well structured, the target is stable, and deployment constraints are strict, a shallow classifier or a small recurrent model may still be the right choice. The goal of this dissertation is not to replace every classical method, but to identify when those methods cease to be adequate.

Third, traditional models help isolate what can be learned from purely quantitative structure. When they perform well, the problem may not require richer

semantic reasoning. When they fail, the failure itself is diagnostic: it indicates that the task depends on context, qualitative meaning, missingness semantics, or transfer properties not captured by the numeric representation.

Fourth, they clarify the nature of the transition made in this dissertation. The move to language models in the next chapter is not a rejection of predictive modeling. It is a response to the specific places where traditional ML/DL breaks. Chapter 3 begins at that point of failure.

2.9 Chapter Summary

This chapter has shown that traditional ML and DL methods provide a necessary starting point for LE forecasting, but they do not provide a sufficient endpoint for the dissertation problem. They work best when the trajectory is short, structured, predominantly numeric, and evaluated within a stable distribution. They become increasingly brittle when the data is contextual, qualitative, sparse, partially observed, and tested under cohort or temporal shift.

The problem, therefore, is not simply that traditional models are smaller or older. The problem is that they are built around assumptions that fit LE data only partially.

These findings primarily address **RQ1** and support **H1**: under small-data, heterogeneous, and shift-prone LE settings, traditional ML/DL baselines remain necessary references but are insufficient as the final representational framework. They can model patterns, but they struggle to interpret behavior. They can ingest sequences, but they do not naturally preserve meaning. They can compensate for missingness numerically, but they do not turn absence into signal. And they can fit a source distribution, but they do not reliably learn what transfers beyond it.

2.10 Bridge to Chapter 3

The next chapter introduces the dissertation's first major response to these limitations: large language models. The transition to LLMs is not merely a change in backbone. It is a change in representational philosophy. Instead of asking a model to predict directly from flattened or numeric trajectories, Chapter 3 asks whether LE data can be verbalized, contextualized, and interpreted as meaningful narratives. That shift begins from the core insight established here: if LE data is fundamentally contextual, then a model pretrained for contextual understanding may be a more appropriate starting point than a model designed only for numeric interpolation.

Chapter 3

Large Language Models (LLMs) for Contextual Modeling of LE Data

3.1 Introduction

Chapter 2 established an important starting point for this dissertation: traditional machine learning and deep learning methods are valuable baselines for LE forecasting, but their performance is fundamentally shaped, and often constrained, by how the data is represented before learning begins. Classical machine learning pipelines rely on engineered features, aggregation, and fixed-vector formulations. Sequence-oriented deep learning models preserve ordering more explicitly, yet they still process LE trajectories primarily as numerical sequences whose meaning is assumed to be stable, directly observable, and largely context-independent. Those assumptions become increasingly brittle as the data becomes more heterogeneous, more qualitative, more incomplete, and more dependent on behavioral interpretation.

This chapter presents the next major shift in the dissertation: the move from primarily numeric modeling to language-based modeling. That shift did not emerge simply because LLMs became prominent in artificial intelligence. It emerged because the research problems studied in this dissertation increasingly

demanded capabilities that conventional modeling pipelines did not naturally provide. LE data is contextual, semantically rich, partially observed, and frequently collected at a scale too small to support robust end-to-end training of large task-specific models. The dissertation therefore turned toward a different strategy: instead of forcing behavioral trajectories into purely numerical representations and learning from scratch, represent them in a form that aligns with the pretraining biases of foundation models.

The central claim of this chapter is that LLMs matter in this dissertation not merely as larger models, but as a different representational paradigm. Once LE trajectories are verbalized appropriately, forecasting ceases to be only a mapping from numbers to labels and becomes a contextual reasoning problem. This reframing enables the model to integrate heterogeneous evidence, exploit semantic regularities, interpret missingness more meaningfully, and generate outputs that reflect richer behavioral understanding than direct classification alone. Across the research arc represented in this chapter, the dissertation progressively demonstrates that: (i) early forecasting can be reformulated as natural language generation; (ii) personalization and contextualization are not optional embellishments but core representational principles; (iii) richer LE datasets make the advantages of language-based modeling more pronounced; (iv) LLMs can support not only prediction but also semantically informed imputation and feature selection; and (v) cross-distribution generalization depends critically on how trajectories are textualized and how outputs are formulated [2, 3, 20, 26, 27, 54].

At the same time, this chapter does not treat LLMs as the end of the story. It shows not only why this methodological transition became necessary, but also which limitations remained after it. The later sections of this chapter therefore show that text-only LE modeling retains a structural weakness: verbalization

improves semantic alignment, but it also linearizes richer temporal structure into a one-dimensional sequence. That limitation becomes the motivation for the multimodal modeling work developed in Chapter 4. Thus, the present chapter plays a dual role in the dissertation. It documents the conceptual and empirical rise of language modeling as the first major answer to the representational failures of numeric LE modeling, and it identifies the remaining bottlenecks that the later multimodal and frozen-prior chapters will address.

To support that argument, the chapter is organized as follows. Section 3.2 explains why LE forecasting naturally pushed the dissertation toward language-based representation. Section 3.3 identifies the properties of LE data that align closely with LLM inductive biases. Section 3.4 formalizes the transformer and adaptation foundations underlying this stage of the research program. Sections 3.6 through 3.10 then trace the methodological evolution across the dissertation's LLM-centered studies: early performance forecasting, contextualized forecasting, broader educational LE modeling, qualitative engagement forecasting, semantically informed handling of missingness, and cross-distribution narrative-based LE generalization. The chapter closes with a synthesis of what the LLM stage contributed to the overall dissertation and a precise explanation of why text-only modeling, despite its gains, is still not sufficient for the final dissertation goal.

3.2 From Numeric Forecasting to Language-Based Representation

The move to language modeling in this dissertation was driven by both empirical dissatisfaction and conceptual necessity. Empirically, the earliest forecasting studies revealed that conventional neural models trained from scratch on small educational datasets did not yield satisfactory performance in low-resource set-

Table 3.1: The LLM stage as an evolving research program rather than a single isolated method.

Study Stage	**Primary Task**	**Main Representational Move**	**Key Dissertation-Level Insight**
Early forecasting with pre-trained LMs	End-of-semester performance prediction from first 2–8 weeks	Reframe forecasting as natural language generation over verbalized student trajectories	The gains come not only from transfer learning, but from aligning the task with a model family already optimized for contextual sequence reasoning [26].
Contextualized language modeling	Early performance prediction with distal, cognitive, and non-cognitive factors	Move from cognitive-only inputs to contextualized behavioral descriptions	Forecasting improves when academic signals are interpreted jointly with learner context rather than in isolation [27].
Broader educational LE modeling	Forecasting from hybrid cognitive, non-cognitive, and background data	Use language as a common representational space for heterogeneous LE modalities	Language models become more useful as LE data becomes more qualitative, hybrid, and semantically difficult for fixed-vector methods [20].
Qualitative engagement forecasting	Weekly engagement prediction under missingness and high-dimensional qualitative inputs	Semantic imputation, feature selection, and text-based forecasting	LLMs can participate in the entire LE pipeline, not just the final classifier, by supporting semantically meaningful preprocessing and modeling [2, 54].
Cross-distribution narrative LE modeling	OOD behavioral forecasting across GLOBEM, LifeSnaps, and MFAFY	Meta-Narratives and Prospective Narrative Generation	Generalization depends on representational design and output formulation, not on model size alone [3].

tings [26]. Conceptually, the shortcomings of those models were not merely matters of insufficient tuning or insufficient depth. They reflected a deeper problem: the dominant modeling pipeline assumed that student trajectories could be represented adequately through numerical sequences and that the learning problem consisted primarily of mapping those sequences to output labels.

That assumption is often too weak for LE data. Let the trajectory for individual i across T time steps be denoted as

$$X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,T}\}, \tag{3.1}$$

where each $x_{i,t}$ may contain cognitive assessments, self-reports, affective states, behavioral indicators, or background-linked context. In traditional forecasting pipelines, the model learns a function

$$\hat{y}_i = f_\theta(X_i), \tag{3.2}$$

with f_θ operating directly on numeric or ordinal inputs. The problem is not that Equation (3.2) is wrong in a formal sense. The problem is that the representation X_i often leaves out what makes LE data difficult and behaviorally meaningful: contextual dependence, heterogeneity of modalities, semantic interpretation of missingness, and uncertainty about what patterns should matter across distributions.

Language-based modeling changes the problem definition by inserting a representation function between the raw trajectory and the model:

$$s_i = \mathcal{V}(X_i; \pi), \tag{3.3}$$

where $\mathcal{V}(\cdot)$ denotes a textualization procedure parameterized by a design policy π that decides what information to include, in what order, at what level of abstraction, and with what contextual framing. The forecasting function then becomes

$$\hat{y}_i = g_\phi(s_i), \tag{3.4}$$

where g_ϕ is a language model or language-based classifier that operates over the verbalized sequence s_i.

This shift is substantial. It transforms the role of representation from a preprocessing afterthought into the core design question. Once LE data is verbalized, the model no longer sees an opaque multivariate time series. It sees a description of a person's evolving state, often phrased in a way that resembles human behavioral interpretation. That makes it possible for the model to leverage pretraining on natural language, to interpret individual features relative to surrounding context, and to treat forecasting as an act of structured reasoning rather than a purely statistical interpolation.

The early forecasting work in this dissertation was the first empirical demonstration of this shift. Student performance prediction was reformulated as a natural language generation problem in which verbalized academic trajectories were used to condition a pre-trained encoder-decoder LM, and the output was another sequence describing predicted end-of-semester performance [26]. From a broader perspective, the importance of that study was not only that it improved results over numeric baselines. It showed that the main methodological bottleneck was not simply learning capacity, but representational compatibility with the task.

3.3 Why LLMs Are a Better Fit for LE Data

The dissertation moves toward LLMs because the structure of LE data aligns naturally with the core strengths of language modeling. Four properties are especially important: context sensitivity, heterogeneous evidence integration, transferability under small-data constraints, and alignment with generative forecasting.

3.3.1 Context sensitivity

A central argument running across this dissertation is that the meaning of behavioral evidence is context dependent. A decline in academic confidence can indicate emerging risk in one student and a transient, adaptive response to challenge in another. Reduced physical activity may reflect depression risk in one context, but ordinary exam-period behavior in another. Traditional fixed-vector models can ingest all these variables, but they do not naturally reinterpret each variable in light of the whole trajectory. LLMs do.

This contextual property mirrors the transition in natural language processing from static to contextual embeddings. Static word vectors assign one representation to a word regardless of use, whereas contextual models represent meaning relative to surrounding text [55, 56]. In an analogous way, language-based LE modeling makes it possible to interpret a behavioral signal relative to the rest of the trajectory, the learner background, and the broader situational context. That is why the move to LLMs should be understood as a move from static interpretation to contextual interpretation, not merely a move from smaller to larger models.

3.3.2 Heterogeneous evidence integration

LE datasets often combine variables that do not live naturally in the same mathematical space: ordinal survey responses, narrative reflections, continuous sensor measures, static background descriptors, and event counts. Numeric modeling pipelines either homogenize them into vectors or require complex fusion architectures. Language provides a different solution: it acts as a common representational interface. Once heterogeneous signals are expressed in text, they can be combined into a single semantically organized prompt without having to force them into the same measurement scale.

This becomes particularly important in educational LE data, where distal background factors, proximal cognitive measures, and proximal non-cognitive measures all contribute to interpretation. In the broader educational dataset used later in this chapter, the trajectory includes 28 repeated non-cognitive dimensions, 41 cognitive measures, and 9 background variables [20]. Language makes it possible to describe these jointly while preserving their differing roles. The background information can appear as context, the repeated performance measures as temporal evidence, and the non-cognitive responses as interpretive cues. That is a more behaviorally faithful representation than flattening them into a single feature vector.

3.3.3 Transfer learning under small-data conditions

Small-data regimes are a recurring theme in this dissertation. LE data is expensive to collect, often requires repeated participation from human subjects, and is therefore rarely available at the scale required to train large specialized models from scratch. The early forecasting datasets in this chapter are small even by

educational modeling standards, and the broader educational LE dataset contains only $N = 48$ students despite its richness [20, 26]. These are not exceptional cases; they are typical of human-centered longitudinal data.

Pre-trained language models offer an attractive solution because they begin with a strong prior acquired from large-scale pretraining [57–59]. Once the trajectory is textualized, the model can reuse knowledge about language, human description, temporal relation, and common-sense behavioral patterns, instead of learning everything from task-specific data alone. In low-resource settings, this prior is not just an efficiency advantage. It is often the difference between a model that can meaningfully generalize and one that merely memorizes the training set.

3.3.4 Alignment with generative forecasting

A further reason LLMs are well matched to LE forecasting is that they enable the task to be framed generatively rather than only discriminatively. In early forecasting, the model generates a textual prediction of future performance rather than directly outputting a class label [26]. Later work pushes this much further by showing that prospective narrative generation better aligns with the contextual reasoning strengths of LLMs than binary classification [3].

From a broader perspective, this is a crucial insight. It means that the gains of language modeling are not only about what goes into the model, but also about what comes out. When the output space is a narrative rather than a hard class boundary, the model can articulate richer relations, softer evidence, and more nuanced behavioral expectations. That design decision becomes central to the generalization results discussed later in the chapter.

3.4 Transformer and Adaptation Foundations

Because the LLM stage is central to the dissertation, it is worth making explicit why transformer-based models, rather than earlier neural architectures, are the foundation of this work. The importance of transformers is not only that they scale well. Their self-attention mechanism makes it possible to compute contextualized representations in which every token can interact directly with every other token in the input sequence [34].

Given an input sequence of token embeddings

$$H^{(0)} = [h_1, h_2, \ldots, h_n], \tag{3.5}$$

self-attention computes learned query, key, and value projections,

$$Q = H^{(0)}W_Q, \qquad K = H^{(0)}W_K, \qquad V = H^{(0)}W_V, \tag{3.6}$$

and produces contextualized outputs via

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{3.7}$$

Multi-head attention extends this idea by learning several attention patterns in parallel, enabling the model to capture different forms of relationship simultaneously.

For LE modeling, the value of this mechanism is conceptual as much as computational. Once the trajectory is verbalized, a statement about confidence in week 2 can be interpreted relative to performance in week 1, background conditions described earlier in the prompt, and engagement changes mentioned later. This ability to form rich intra-sequence relationships is precisely what makes

transformers suitable for contextual behavioral interpretation.

The dissertation's early LLM work relied on encoder-decoder models such as FLAN-T5, which are natural fits for conditional generation tasks [57]. Later work used both encoder-only and decoder-only models depending on the task, revealing that architecture interacts with data characteristics. For sparse qualitative classification, encoder-style models sometimes perform better because their bidirectional token encoding is effective for sentence-level classification, whereas decoder-only models are often better suited to open-ended generation [2, 54]. This architectural nuance matters because it reinforces a recurring argument in this dissertation: the success of foundation models depends on matching model behavior to representational demands, not on using the largest available model indiscriminately.

3.4.1 Generative and discriminative training objectives

When forecasting is formulated as natural language generation, the model is adapted with the standard teacher-forced sequence objective:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T_y} \log p_\theta(y_t \mid y_{<t}, s_i), \tag{3.8}$$

where s_i is the textualized input trajectory and y_t is the t-th target token of the generated forecast. This objective was central to the earliest LLM forecasting work in the dissertation and returns later in Prospective Narrative Generation [3, 26].

When the task is formulated as direct classification, the adapted representation is mapped to a label distribution and optimized via cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\sum_{c \in \mathcal{Y}} \mathbb{1}[y = c] \log p_\theta(c \mid s_i). \tag{3.9}$$

The dissertation's later work explicitly compares these two formulations and shows that the generative option is often better aligned with LLM inductive biases for generalization [3].

3.4.2 Parameter-efficient adaptation

As the studies in this chapter became more ambitious, parameter-efficient fine-tuning also became important. The dissertation used techniques such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) to adapt large pretrained models without updating every parameter [41,42]. In LoRA, a pretrained weight matrix W_0 is adapted through a low-rank update:

$$W = W_0 + \Delta W = W_0 + BA, \tag{3.10}$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with rank $r \ll \min(d, k)$. The practical significance of this equation in the dissertation is twofold. First, it made experimentation with larger models feasible under realistic computational constraints. Second, it reduced the temptation to interpret all improvements as consequences of brute-force adaptation; instead, the chapter repeatedly shows that representational design remains the dominant factor.

3.5 Verbalization as Representational Design

The dissertation's LLM stage is not just about swapping one model family for another. It is about learning how to transform LE trajectories into textual forms that preserve the right information. Verbalization is therefore not a cosmetic conversion from numbers to words. It is a representational design choice that determines what the model can reason about.

A generic LE textualization can be expressed as

$$s_i = \mathcal{T}(b_i, c_{i,1:T}, n_{i,1:T}, m_{i,1:T}; \rho), \tag{3.11}$$

where b_i denotes background variables, $c_{i,1:T}$ cognitive measures, $n_{i,1:T}$ non-cognitive measures, $m_{i,1:T}$ missingness or metadata, and ρ denotes a policy controlling granularity, wording, temporal markers, and task instructions. The studies in this chapter effectively investigate different settings of ρ.

Three design questions recur across the work:

First, *what level of detail should be preserved?* A fully verbalized sequence can keep a large amount of local information, but may become long, noisy, and difficult to process. A summarized description reduces length, but risks discarding weak but important temporal cues.

Second, *how should context be expressed?* Distal background factors and non-cognitive states can be appended as raw variables, but the research in this chapter shows that they are more useful when integrated as part of a coherent contextual description.

Third, *how should absence or missingness be described?* In LE data, a missing response is often behaviorally meaningful. Treating it as zero or silently imputing it can erase information. Language-based descriptors make it possible to encode that absence in semantically interpretable ways.

The methodological arc of this chapter can therefore be read as a progressive refinement of textualization policy ρ. Early work uses explicit verbalization with instructions. Later work adds richer contextual framing, temporal cues, descriptors for missingness, and eventually narrative summarization. ConText-LE makes this representational perspective fully explicit by comparing multiple input tex-

tualizations and showing that generalization depends strongly on which one is chosen [3].

3.5.1 Textual representation families before and within ConText-LE

The work in this chapter progressively explores several families of textual representation, even before they are named explicitly in ConText-LE. At one end are near-literal verbalizations that preserve local measurements and explicit temporal order. At the other end are synthetic narratives that emphasize pattern interpretation over exhaustive detail. Both extremes have value. Literal encodings preserve evidence granularity and support faithful recovery of rare events, whereas narrative encodings expose higher-level regularities and behavioral themes that may be easier for an LLM to reason over under distribution shift.

This tension is important because it reframes a practical modeling choice as a dissertation-level design trade-off. A representation can be evaluated not only by how much information it contains, but also by what kind of reasoning it invites from the model. The chapter's early forecasting studies implicitly favor detailed verbalization because their main challenge is learning from very little data without discarding evidence. ConText-LE later shows that once the objective becomes cross-distribution generalization, a more interpretive representation can be preferable because it suppresses source-specific noise and foregrounds transferable behavioral structure.

Table 3.2 summarizes this progression.

A related issue is task scaffolding. The early studies in this chapter consistently found that LLM performance improved when input prompts explicitly stated the forecasting goal, organized the evidence by component, and used natural instructional framing rather than raw data dumps. This is consistent with the

Table 3.2: Textual representation families explored across the LLM stage of the dissertation.

Representation Family	What It Preserves Well	Main Strength	Main Risk or Limitation
Literal verbalized sequence	Local measurements, explicit order, exact feature values	Faithful to the observed trajectory; useful when fine-grained evidence matters	Long prompts, noisy inputs, and weak abstraction over source-specific details
Structured value listing / language string	Feature-wise chronology in a compact, regular format	Easier for the model to parse repeated values and missingness markers consistently	Still linearizes the data and may underemphasize cross-feature interpretation
Statistical summary	Global aggregates such as mean, range, and variability	Efficient compression and robustness to prompt length constraints	Can suppress temporal transitions, change points, and local anomalies
Contextual narrative / Meta-Narrative	Salient trends, interactions, and behavioral interpretation	Best aligned with LLM reasoning and transfer across contexts	May omit weak but predictive details and depends on high-quality summarization

broader literature on instruction tuning and text-to-text transfer [57–59]. In LE modeling, instructions serve an especially important role because they tell the model what kind of future-oriented reasoning is expected from the input sequence. Without that scaffolding, the verbalized trajectory can remain only a description. With it, the sequence becomes an evidential prompt for behavioral forecasting.

3.6 Early Forecasting as Natural Language Generation

The first major empirical step in this dissertation's LLM trajectory was the reformulation of early performance forecasting as a natural language generation problem [26]. This work examined the challenging setting of forecasting end-of-semester student performance using only the earliest part of a semester-long trajectory. The task is practically important because early-warning systems are most

useful when predictions are available soon enough to support timely intervention, yet it is methodologically difficult because only limited evidence is available in the first few weeks.

3.6.1 Task setting and data structure

The study used data from $N = 48$ first-year college students in an introductory programming course. The input combined three categories of information: distal background factors, proximal cognitive data, and proximal non-cognitive data. After feature selection to fit model context limits, the verbalized representation still preserved a meaningful cross-section of the academic trajectory, including background, assessment history, and repeated engagement-related measures [26].

Forecasting was evaluated at three early points in a 16-week semester: after 2 weeks, 4 weeks, and 8 weeks. Instead of directly training a numeric classifier, the work fine-tuned FLAN-T5 models of varying capacity to generate a textual prediction corresponding to one of four end-of-semester performance groups. Importantly, the numeric data was first verbalized and then augmented to reduce class imbalance. This combination of textualization, instruction-style prompting, and augmentation defined the dissertation's first working recipe for adapting LLMs to LE data.

3.6.2 Why this stage mattered

This study was important far beyond its immediate benchmark result. First, it showed that early forecasting in low-resource educational settings could benefit from transfer learning even when no task-specific pretrained numeric model existed. Second, it demonstrated that the useful unit of transfer was not a shared feature space with another educational prediction task, but the language prior

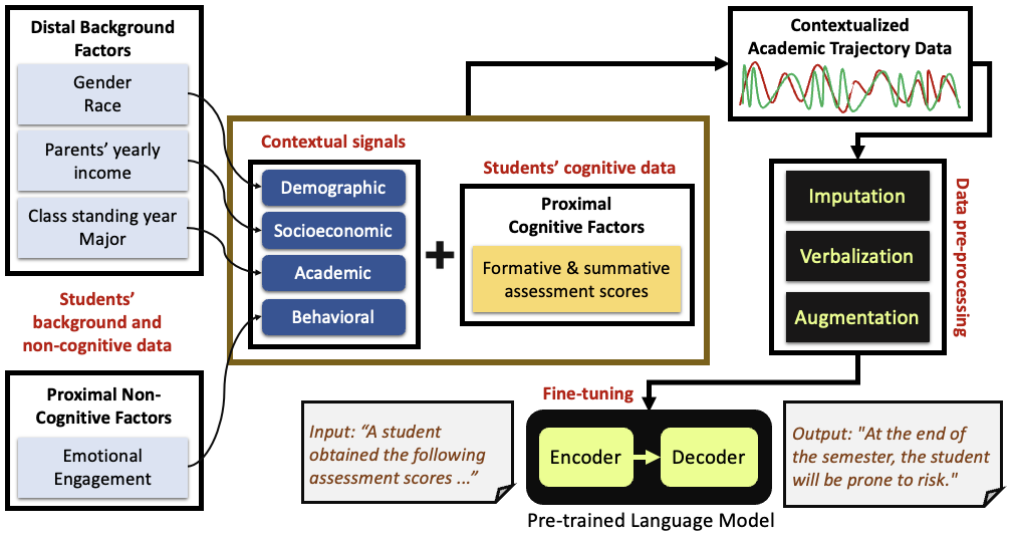


Figure 3.1: Early LLM-based forecasting pipeline adapted from the dissertation's initial language-modeling work. Numeric longitudinal trajectories are verbalized, contextualized, augmented, and used to fine-tune a pre-trained model for sequence generation of future performance [26, 27].

of a general-purpose model. Third, it introduced two ideas that became durable throughout the dissertation: *personalization* through distal features and *contextualization* through non-cognitive signals.

The results were strong enough to justify the shift. The best fully contextualized FLAN-T5-Large model forecasted student performance with 77% accuracy by the end of week 2 and 89% accuracy by the 8-week setting, while also achieving perfect recall for the at-risk group at critical horizons [26]. From the perspective of the dissertation, the significance of these results lies not only in the absolute numbers, but in what they imply. A model pretrained on text was better able to exploit scarce educational data once that data was represented as language and enriched with context.

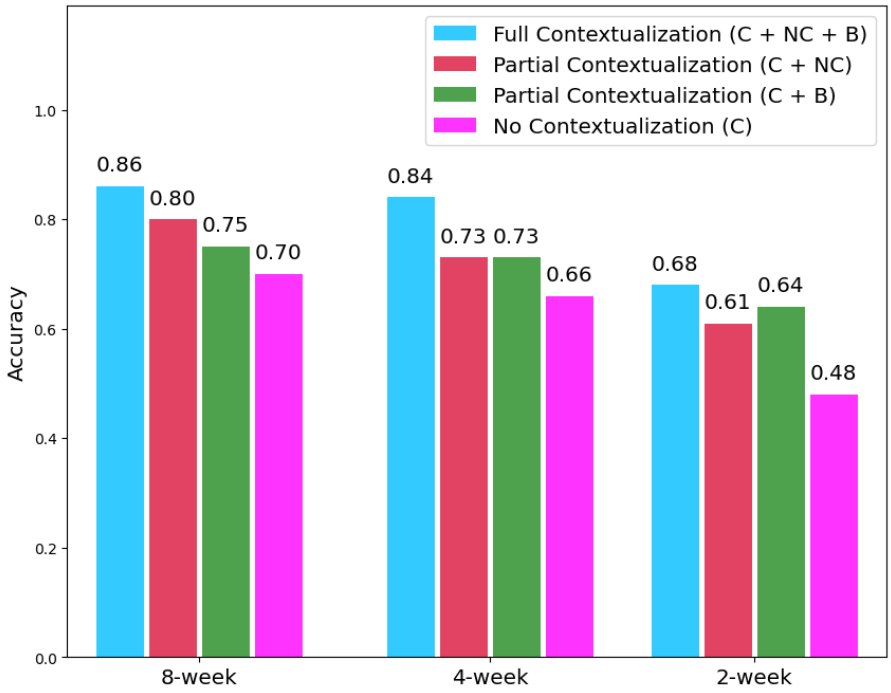


Figure 3.2: Representative early forecasting comparison from the dissertation's initial LLM experiments. The most reliable gains emerge when cognitive signals are combined with contextual and background information rather than modeled in isolation [26, 27].

3.6.3 Personalization and contextualization as dissertation principles

The conceptual importance of the early forecasting study lies in how it redefined what the model should pay attention to. Distal factors were used to introduce personalization: the hypothesis that background conditions shape how a student's trajectory should be interpreted. Non-cognitive measures were used to introduce contextualization: the hypothesis that motivation, engagement, and related experiential variables provide explanatory signals that pure performance histories miss. These were not merely add-on features. They were the first explicit argument in the dissertation that future prediction from human-centered longitudinal data requires representation of the surrounding context, not only the apparent target signal.

In retrospect, this stage became the conceptual entry point for the whole dissertation. It showed that the dissertation's later emphasis on richer representations, contextual meaning, and nontrivial output design had already been foreshadowed in the earliest educational forecasting problem.

3.7 Contextualized Forecasting as a Modeling Principle

The next major step was to clarify that contextualization was not incidental to performance gains, but a core modeling principle [27]. In this stage, the research focus shifted from simply asking whether an LM can outperform numeric baselines to asking *why* contextualized representations work better.

The contextualized forecasting framework explicitly organized the student's state into distal, proximal cognitive, and proximal non-cognitive components. That structure made a dissertation-level claim possible: predictive signals do not exist independently of the interpretive frame in which they are embedded. A weak quiz score, for example, means something different when accompanied by improving motivation than when accompanied by declining engagement and low confidence. The contextualized representation let the model reason about those joint patterns.

This insight is crucial because it transformed a successful empirical trick into a general methodological argument. If behavior is situated, then behavioral forecasting must also be situated. The representation should therefore not merely list features but should present them as an evolving portrait of the learner's state. That argument later generalizes beyond education to mental health and behavior forecasting in cross-distribution settings.

Formally, the contextualized representation used in this stage can be under-

stood as a structured prompt template,

$$s_i = \mathcal{T}(D_i, C_i^{1:t}, NC_i^{1:t}), \tag{3.12}$$

where D_i contains distal information, $C_i^{1:t}$ the cognitive history up to time t, and $NC_i^{1:t}$ the non-cognitive history up to the same time. The important point is not just that all three are concatenated, but that they are presented as semantically interpretable components of one narrative input.

From the dissertation's perspective, this stage established a principle that survives across all subsequent chapters: when the target phenomenon is human behavior, context is not a nuisance variable. It is part of the signal.

3.8 Expansion to Broader LE data

Once the dissertation moved beyond narrow early performance forecasting, the representational challenge became more demanding. The broader educational LE dataset used in the next stage combined static background data, time-varying cognitive measurements, and a large set of repeated qualitative non-cognitive observations [20]. This was no longer a small extension of the early forecasting setting. It was a transition into a more faithful approximation of real LE data.

3.8.1 Why the richer dataset mattered

The dataset contained 28 non-cognitive dimensions, 41 quantitative cognitive measures, and 9 background dimensions. It was small in sample size but high in semantic richness. It also contained substantial missingness, heterogeneous measurement types, and imperfect temporal alignment between cognitive and non-cognitive sources [20]. These properties made it an excellent stress test for the

dissertation's emerging claim that representation, rather than raw model size, is the key challenge.

A concise summary of the main datasets used across the chapter is given in Table 3.3. The progression from the early educational setting to cross-distribution behavioral datasets makes visible how the representational burden grows as the dissertation advances.

3.8.2 Data enrichment, temporal markers, and missingness descriptors

The richer educational LE setting made it clear that verbalization alone was not enough. The textual representation had to be carefully designed to cope with long sequences, sparse observations, and temporal ordering. This stage therefore introduced a more sophisticated input pipeline involving verbalization, augmentation, explicit task instructions, and contextual cues about temporal position [20].

One of the most informative findings at this stage concerned temporal reasoning. Experiments with randomized weekly and daily orderings showed that the models were learning from temporal structure, but only imperfectly. Full temporal randomization caused substantial performance drops, indicating that the models were not merely exploiting static marginal statistics. Yet pseudo-randomization experiments also suggested that explicit temporal tags alone were not always enough for the model to reconstruct the most meaningful temporal relations [20]. This result became conceptually important later in the dissertation, because it hinted that text can capture temporal semantics but may still struggle to preserve temporal structure faithfully.

Missingness also emerged as a representational issue rather than a preprocessing detail. Replacing missing responses with contextually relevant descriptors such

Table 3.3: Representative datasets and task settings used across the LLM stage of the dissertation.

Dataset / Study	Scale	Input Composition	Task	Why It Matters for the Dissertation
Early programming-course forecasting	48 students	Distal background, proximal cognitive, proximal non-cognitive	End-of-semester performance forecasting from 2, 4, and 8 weeks	First evidence that LE forecasting can be reframed as language generation under small-data constraints [26, 27].
Broader educational LE dataset (MFAFY-style educational setting)	48 students; 78 feature dimensions before selection	28 non-cognitive repeated measures, 41 cognitive measures, 9 background features	Early academic performance forecasting from hybrid LE data	Shows that richer, more heterogeneous LE data increases the advantage of language-based representation [20].
Qualitative engagement forecasting cohort	96 students; 960 trajectories	High-dimensional qualitative non-cognitive data plus background information	Weekly engagement forecasting across four constructs	Demonstrates that LLMs can support semantic preprocessing, imputation, and feature selection in addition to prediction [2, 54].
GLOBEM	661 participants over 4 years	Mobile sensing, behavioral activity, sleep, communication, mood	Cross-temporal depression-risk forecasting	Establishes the dissertation's most demanding OOD benchmark for behavioral generalization [3, 16].
LifeSnaps	39 participants over 4 months	Physiological signals, activity, and self-reports	Cross-temporal anxiety forecasting	Tests whether narrative-based LE modeling extends beyond education to smaller multimodal health data [3, 28].
MFAFY (cross-year setting)	96 students across two academic years	Qualitative educational LE trajectories	Cross-year engagement forecasting	Provides a text-dominant OOD benchmark and connects the educational work to the dissertation's generalization theme [3].

as "Skipped the question" produced stronger results than generic markers such as "N/A", showing that LLMs are sensitive to the semantics of how absence is described [20]. This finding is more important than it first appears. It reveals that language models do not merely tolerate missingness; they interpret it. That insight later becomes foundational to the dissertation's broader treatment of missingness as information rather than noise.

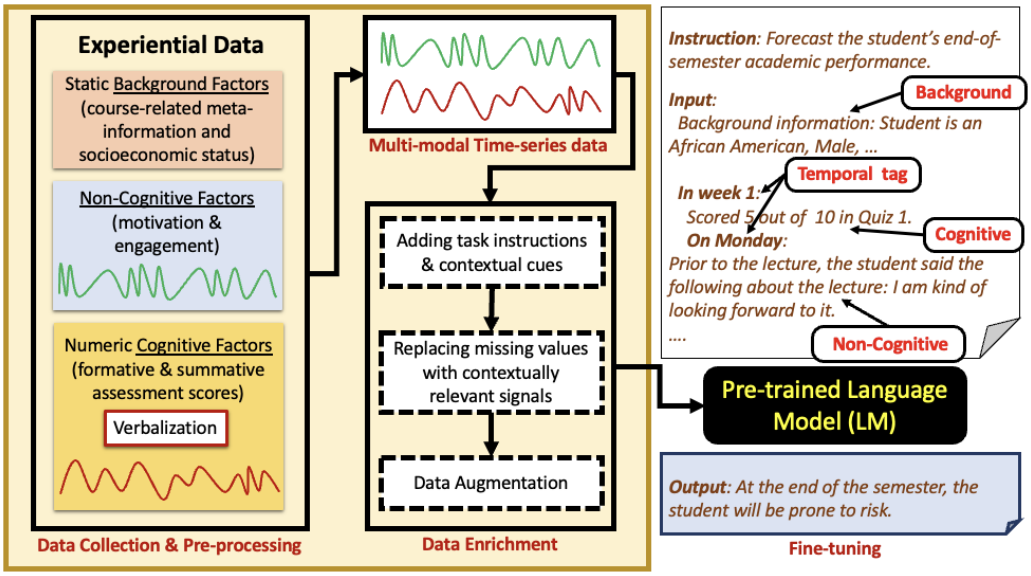


Figure 3.3: Pipeline from the broader educational LE modeling stage. The key transition is from raw hybrid LE data to enriched language sequences that incorporate missingness handling, augmentation, instructions, and temporal cues [20].

3.8.3 What this stage revealed

The main contribution of this stage was not simply that another LM benchmark improved. It clarified that language-based LE modeling succeeds only when representation design does a large share of the work. The experiments showed that: (i) experiential non-cognitive data can be more predictive than cognitive-only signals when encoded effectively; (ii) correlations across modalities matter; (iii) explicit temporal and contextual wording influence model behavior; and

(iv) semantically appropriate treatment of missingness can materially improve results [20]. These lessons were later absorbed directly into the dissertation's more advanced frameworks.

3.9 LLMs for Qualitative Engagement Forecasting

The next phase of the research program moved into an even more demanding regime: forecasting weekly engagement-related outcomes from high-dimensional qualitative longitudinal data with substantial missingness, class imbalance, and semantically entangled self-reports. This phase is particularly important in the dissertation because it forced the language-modeling argument to confront one of the hardest versions of LE data: sparse weekly non-cognitive trajectories whose predictive signal is distributed across subjectively phrased responses, missing answers, and background context rather than in dense numeric measurements. In this setting, a model cannot succeed by merely tracking surface statistics. It must infer what the qualitative responses suggest about the student's evolving engagement state, how missingness should be interpreted, and which dimensions are worth attending to [2, 54].

The setting matters because the target variables are not coarse downstream labels alone. They are nuanced educational constructs such as Lecture Engagement Disposition (LED), Academic Self-Efficacy (ASE), Performance Self-Evaluation (PSE), and Academic Identity and Value Perception (AIVP). These constructs are interpretive, weakly aligned with any single observable indicator, and often collected through repeated self-reports whose absence can itself be informative. This is precisely the kind of task in which the dissertation's representation-first view is most demanding. If language-based modeling genuinely offers an advantage,

that advantage should become most visible when the data is sparse, qualitative, incomplete, and behaviorally meaningful. The results in this phase show that it does.

3.9.1 Why this setting was methodologically different

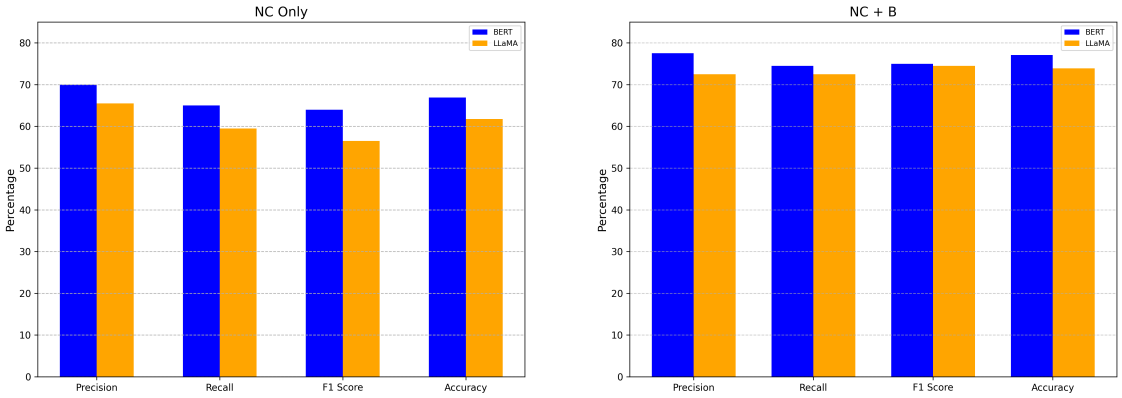
The qualitative engagement setting differs from the earlier stages of the chapter in three important ways. First, the inputs are predominantly non-cognitive rather than assessment-based. Second, the dimensionality of candidate features is high relative to the amount of data. Third, missingness is no longer a side issue. It is one of the defining structural properties of the data. These three properties force the modeling pipeline to answer questions that earlier performance-forecasting settings could postpone: what should be done with missing responses, which qualitative dimensions should be verbalized, and which language-model architectures are best suited to discriminative classification over sparse self-report sequences [2, 54].

The dissertation addresses these questions in a staged manner. The first stage establishes that language models remain useful even when the inputs are limited to qualitative non-cognitive variables. The second stage shows that background information materially improves forecasting because it provides context that helps disambiguate subjective responses. The third stage pushes the pipeline further by treating imputation, feature selection, and forecasting as semantically related components of one LLM-centered process rather than three disconnected preprocessing steps.

3.9.2 Encoder-only versus decoder-only evidence

The lecture-engagement experiments provided an important architectural comparison [54]. The comparison was between BERT-base, an encoder-only bidi-

rectional transformer pretrained with masked-language modeling [56, 60], and Llama 3.1 8B, a decoder-only autoregressive transformer from the Llama family [39]. Using only non-cognitive inputs, BERT-base outperformed Llama 3.1 8B and delivered stronger precision-recall balance for weekly engagement classification [54]. When background variables were appended to the same qualitative trajectories, the performance of both model families improved, but the encoder-style model retained a clear advantage. This is an important corrective to simplistic claims that larger decoder-only models dominate every LE task. For sparse qualitative classification, encoder-only models provide an inductive bias that is often better matched to the task: they encode the entire input sequence bidirectionally, support fine-grained classification decisions, and are less dependent on generative behavior to produce a stable output [54, 56, 60].



(a) Non-cognitive inputs only.

(b) Non-cognitive plus background inputs.

Figure 3.4: Representative encoder-only versus decoder-only evidence from the lecture-engagement forecasting stage. Background context consistently strengthens performance, and encoder-only modeling remains particularly effective for sparse qualitative classification [54].

The broader importance of this comparison is conceptual. It shows that the LLM stage of the dissertation is not reducible to scaling up a single architecture family. Instead, the dissertation treats language-based modeling as a representa-

tional paradigm whose downstream realization may take different architectural forms depending on the task. Where the goal is narrative generation, decoder-only instruction-tuned models become natural. Where the goal is binary classification over short, sparse qualitative trajectories, encoder-style transformers can be more appropriate.

3.9.3 Missingness, context, and reliable representation

At this point in the research program, missingness moved from being a nuisance to being a core representational problem. In many conventional pipelines, missing values are repaired before modeling through deletion, interpolation, mean substitution, or model-based imputation [61–65]. That treatment is often serviceable when missingness is light and approximately random. Qualitative LE data, however, frequently violates those assumptions. A skipped weekly reflection, an unanswered self-efficacy item, or an omitted belonging-related response may reflect disengagement, fatigue, uncertainty, avoidance, or contextual instability. When that is the case, the absence of a value is itself part of the behavioral trajectory.

For each value $x_{i,t,j}$, let the missingness indicator be defined as

$$m_{i,t,j} = \begin{cases} 1, & \text{if } x_{i,t,j} \text{ is observed,} \\ 0, & \text{if } x_{i,t,j} \text{ is missing.} \end{cases} \tag{3.13}$$

Let X_{obs} and X_{mis} denote the observed and missing portions of a trajectory. Then the missingness process is characterized by

$$P(M \mid X_{\text{obs}}, X_{\text{mis}}), \tag{3.14}$$

and the classical taxonomy distinguishes three regimes [61, 62, 66]:

$$P(M \mid X_{\text{obs}}, X_{\text{mis}}) = P(M) \qquad \text{(MCAR)}, \tag{3.15}$$

$$P(M \mid X_{\text{obs}}, X_{\text{mis}}) = P(M \mid X_{\text{obs}}) \qquad \text{(MAR)}, \tag{3.16}$$

$$P(M \mid X_{\text{obs}}, X_{\text{mis}}) \neq P(M \mid X_{\text{obs}}) \qquad \text{(MNAR)}. \tag{3.17}$$

The dissertation's argument is that qualitative LE forecasting often behaves much closer to the MNAR setting than to MCAR. The important design question is therefore not only how to estimate the missing value, but how to represent the fact and possible meaning of the missingness event. In ordinary imputation, one seeks

$$\hat{x}_{i,t,j} = \Gamma(X_{\text{obs}}, M), \tag{3.18}$$

where $\Gamma(\cdot)$ returns a best-guess scalar value. In the language-based formulation developed here, the objective is instead to build a representation

$$r_{i,t,j} = \Phi(x_{i,t,j}, m_{i,t,j}, \mathcal{C}_{i,t}), \tag{3.19}$$

that preserves both the observed value when available and the contextual meaning of absence when it is not. When $x_{i,t,j}$ is missing, the representation can take the form of a descriptor

$$d_{i,t,j} = \Psi(X_{i,\text{obs}}, M_i, \mathcal{C}_{i,t}), \tag{3.20}$$

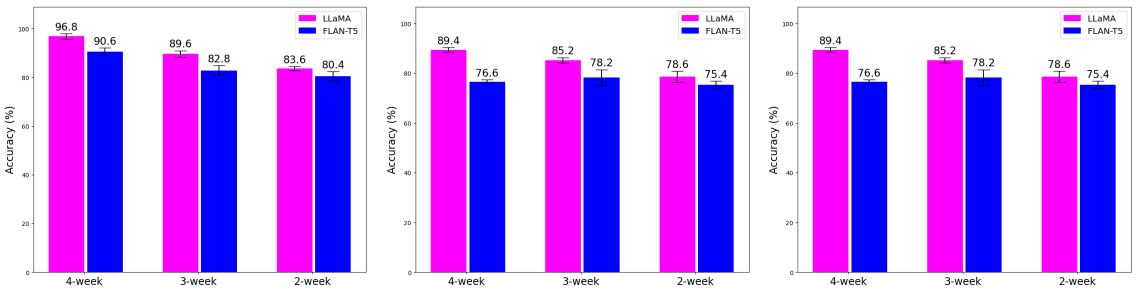
where $\Psi(\cdot)$ is a language-based generator conditioned on surrounding trajectory evidence and local context.

This distinction is not cosmetic. It changes the semantics of the downstream input. A generic marker such as *N/A* tells the model only that the value is absent.

A contextually meaningful phrase such as *Skipped the question* makes the absence interpretable as part of a human interaction process. The broader theory developed later in the dissertation is that reliable LE modeling begins at exactly this level of representational choice: if the model receives a semantically flattened version of the trajectory, later predictive sophistication cannot fully recover the information that was lost.

3.9.4 Early evidence that wording matters for missingness

The experiments provided the first direct evidence that wording matters for missingness representation [20]. In this setting, missing experiential responses were replaced using three different descriptors: a contextually meaningful phrase (*Skipped the question*), a generic but semantically plausible marker (*N/A*), and an intentionally contextually incorrect placeholder (*Hello, World!*). The purpose was not only to recover performance under incomplete data, but also to test whether language models distinguish among different forms of absence description.



(a) *Skipped the question* (b) *N/A* (c) *Hello, World!*

Figure 3.5: Investigation of missing-value wording in the broader educational LE modeling stage. Contextually meaningful descriptors led to the strongest downstream forecasting performance, showing that LMs are sensitive to the semantics of how missingness is verbalized [20].

The results were revealing. Replacing missing values with *Skipped the question* produced the highest performance for both LLaMA and FLAN-T5. Replacing them

with *N/A* caused a notable decline, including a drop of 7.4 percentage points for the 4-week LLaMA setting and about 14 points for FLAN-T5 in the corresponding experiment [20]. The *Hello, World!* condition also reduced performance, but the degradation was smaller than might be expected if missingness were semantically inert. The important conclusion is not that one phrase is universally optimal. It is that the models interpret the descriptor itself as part of the behavioral evidence. This became the empirical seed for the dissertation's later claim that missingness in LE data is often better handled as semantic context than as mere numerical repair.

3.9.5 CRILM: from local intuition to a general missingness-aware framework

The next step in the dissertation was to turn this local educational observation into a general framework. That framework is **CRILM**, a context-aware approach for enhancing data imputation with pre-trained language models [1]. CRILM is important in the dissertation because it makes the missingness argument fully explicit and tests it under controlled missingness mechanisms rather than only in a single educational dataset.

The central move in CRILM is to stop treating imputation as purely numeric estimation and instead cast missingness handling as a language-centered representational problem. Consider a sample-level input vector

$$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d}), \tag{3.21}$$

with missingness mask

$$m_i = (m_{i,1}, m_{i,2}, \ldots, m_{i,d}). \tag{3.22}$$

For each feature j, CRILM defines a transformed entry

$$\tilde{x}_{i,j} = \begin{cases} x_{i,j}, & \text{if } m_{i,j} = 1, \\ d_{i,j}, & \text{if } m_{i,j} = 0, \end{cases} \tag{3.23}$$

where $d_{i,j}$ is a feature-specific textual descriptor rather than a scalar estimate. The resulting missingness-aware record is then verbalized into a contextual sequence

$$z_i = \mathcal{V}(\tilde{x}_{i,1}, \tilde{x}_{i,2}, \ldots, \tilde{x}_{i,d}; \tau), \tag{3.24}$$

where $\mathcal{V}$ is a verbalization operator and τ encodes template decisions such as feature wording, order, and task instructions. A downstream language model then performs prediction directly from the missingness-aware text:

$$\hat{y}_i = f_\theta(z_i). \tag{3.25}$$

This creates a dual-phase design. In the first phase, a stronger conversational language model is used to generate semantically appropriate descriptors for missing features. In the second phase, a smaller pre-trained model is fine-tuned on the resulting missingness-aware textual data for the downstream task. Conceptually, the first phase contributes semantic prior knowledge, while the second phase contributes efficient task adaptation. This division is important for the dissertation because it anticipates the later logic of using strong frozen or semi-frozen priors to regularize downstream learning rather than relying only on unrestricted end-to-end fitting.

The deeper conceptual significance of CRILM lies in what it refuses to assume. Most classical imputers seek a plausible numeric value under assumptions about

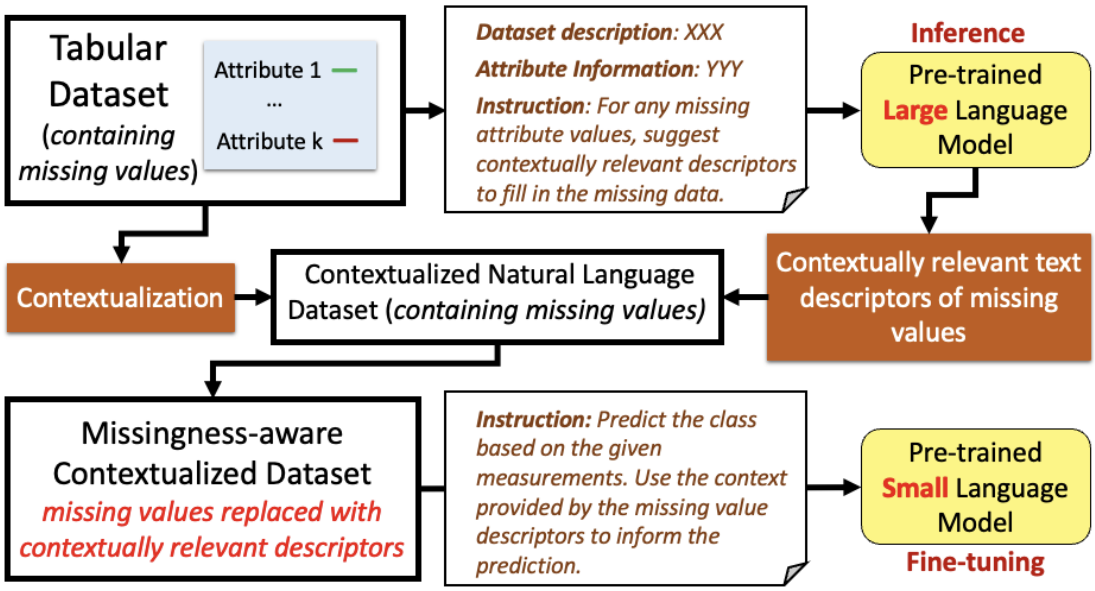


Figure 3.6: Overview of CRILM. Missing values are replaced by contextually relevant descriptors, verbalized into a missingness-aware language sequence, and then used to fine-tune a smaller downstream language model [1].

the data distribution, often relying on the idea that the missingness pattern is MCAR or MAR, or at least tractable with distributional modeling. CRILM instead asks whether a language model can exploit prior knowledge and contextual cues to preserve the *meaning* of absence even when estimating the exact latent scalar is difficult, biased, or arguably the wrong objective. In this sense, CRILM is less a competitor to every imputation algorithm in principle than a challenge to the assumption that all missingness problems should first be reduced to scalar reconstruction.

3.9.6 Controlled evaluation across MCAR, MAR, and MNAR

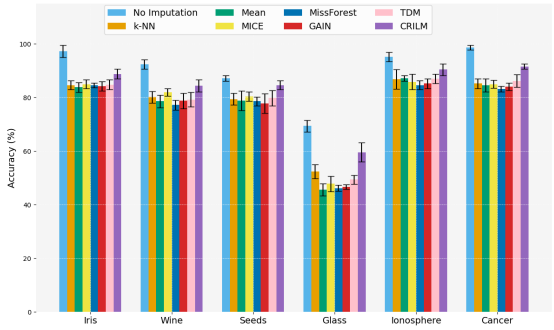
To test that claim rigorously, the CRILM study constructed controlled missingness scenarios across six UCI classification datasets and evaluated performance under three standard regimes: MCAR, MAR, and MNAR [1]. Up to 30% missing values were introduced synthetically so that the effect of the missingness mech-

anism could be isolated rather than confounded with unrelated data-collection artifacts. CRILM was then compared against a diverse set of numeric baselines, including mean imputation, k-NN, MissForest, MICE, GAIN, and transformed distribution matching. The downstream models were a decoder-only LLaMA 2 model and an encoder-decoder FLAN-T5 model.

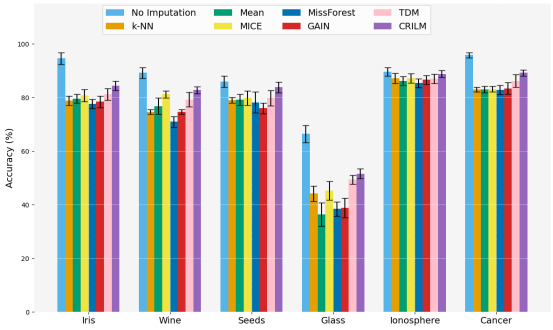
The importance of this design for the dissertation is methodological. Earlier educational experiments established that descriptor wording matters. CRILM shows that the same principle survives when the evaluation is broadened beyond one domain and confronted with adversarially difficult missingness regimes. It therefore supplies the missing bridge between a local LE observation and a general representational claim.

Several detailed results matter for the dissertation's argument. Under MCAR, CRILM remained competitive or superior even though the missingness process was maximally unstructured, suggesting that semantic replacement does not depend on a favorable causal story to be useful. Under MAR, the margins narrowed in some cases because classical methods can partially exploit correlations among observed variables, but CRILM still retained an edge. Under MNAR, where the missingness process is most misaligned with naive scalar recovery, the advantage became especially clear. This is precisely the regime that most resembles many real LE settings, where missing responses often reflect the latent state of the participant rather than mere observation noise.

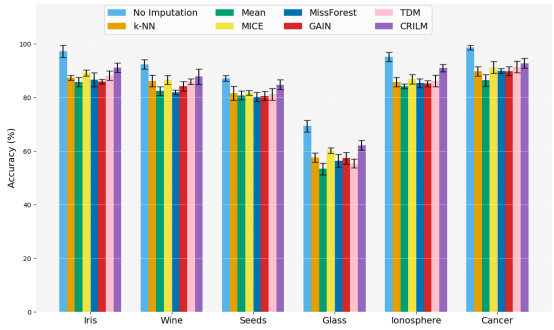
Tables 3.4–3.6 report the challenge-dataset analysis that is most relevant for the dissertation. For LLaMA, CRILM raised Glass accuracy from 52.4% to 59.6% under MCAR, from 60.2% to 62.2% under MAR, and from 44.8% to 54.8% under MNAR. On Seeds, the corresponding gains were 80.4% to 84.6%, 81.8% to 84.8%, and 76.4% to 82.4%. On Wine, CRILM improved 82.0% to 84.4%, 86.6% to 87.8%, and 75.6%



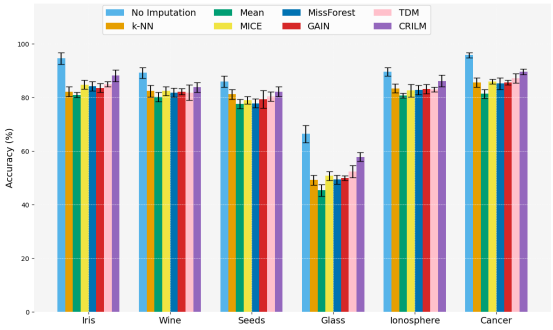
(a) MCAR with LLaMA



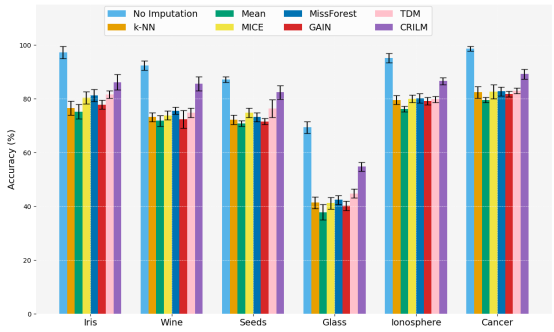
(b) MCAR with FLAN-T5



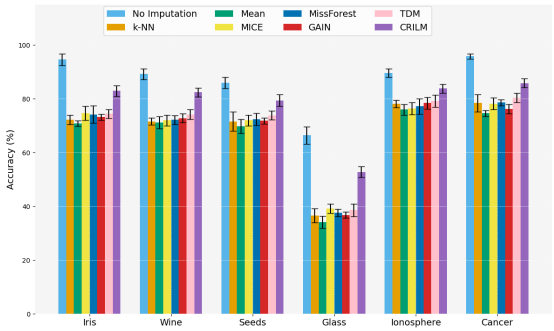
(c) MAR with LLaMA



(d) MAR with FLAN-T5



(e) MNAR with LLaMA



(f) MNAR with FLAN-T5

Figure 3.7: CRILM compared with established numeric imputation baselines across MCAR, MAR, and MNAR missingness patterns using both LLaMA and FLAN-T5 for downstream prediction [1].

to 85.6%. FLAN-T5 showed the same pattern: on Glass, CRILM improved 45.6% to 51.6% under MCAR, 52.4% to 57.8% under MAR, and 39.2% to 52.8% under MNAR; on Seeds, it improved 79.8% to 83.8%, 81.2% to 82.2%, and 73.8% to 79.4%; and on Wine, it improved 81.2% to 82.4%, 82.4% to 83.8%, and 74.2% to 82.4% [1]. These are not marginal fluctuations. They show that the representational treatment of missingness measurably changes downstream discriminative performance across distinct mechanisms and model families.

Table 3.4: Detailed CRILM gains over the best numeric baseline on three challenging datasets under MCAR missingness [1].

Dataset	LM	Best Baseline	CRILM	Gain
Glass	LLaMA	52.4 (k-NN)	59.6	+7.2
Glass	FLAN-T5	45.6 (TDM)	51.6	+6.0
Seeds	LLaMA	80.4 (MICE)	84.6	+4.2
Seeds	FLAN-T5	79.8 (MICE)	83.8	+4.0
Wine	LLaMA	82.0 (MICE)	84.4	+2.4
Wine	FLAN-T5	81.2 (MICE)	82.4	+1.2

Table 3.5: Detailed CRILM gains over the best numeric baseline on three challenging datasets under MAR missingness [1].

Dataset	LM	Best Baseline	CRILM	Gain
Glass	LLaMA	60.2 (MICE)	62.2	+2.0
Glass	FLAN-T5	52.4 (TDM)	57.8	+5.4
Seeds	LLaMA	81.8 (MICE)	84.8	+3.0
Seeds	FLAN-T5	81.2 (k-NN)	82.2	+1.0
Wine	LLaMA	86.6 (MICE)	87.8	+1.2
Wine	FLAN-T5	82.4 (k-NN)	83.8	+1.4

Table 3.6: Detailed CRILM gains over the best numeric baseline on three challenging datasets under MNAR missingness [1].

Dataset	LM	Best Baseline	CRILM	Gain
Glass	LLaMA	44.8 (TDM)	54.8	+10.0
Glass	FLAN-T5	39.2 (MICE)	52.8	+13.6
Seeds	LLaMA	76.4 (TDM)	82.4	+6.0
Seeds	FLAN-T5	73.8 (TDM)	79.4	+5.6
Wine	LLaMA	75.6 (MissForest)	85.6	+10.0
Wine	FLAN-T5	74.2 (TDM)	82.4	+8.2

Two points are especially important. First, the gains are largest in the most difficult settings. This strengthens the dissertation's claim that semantically informed missingness handling is not a cosmetic intervention that matters only on easy data. Second, the gains appear across both a larger decoder-only model and a smaller encoder-decoder model. The representational idea therefore transfers across model classes, which supports the dissertation-level argument that the improvement comes from the input design rather than from one specific architecture.

3.9.7 Feature-specific descriptors, not generic placeholders

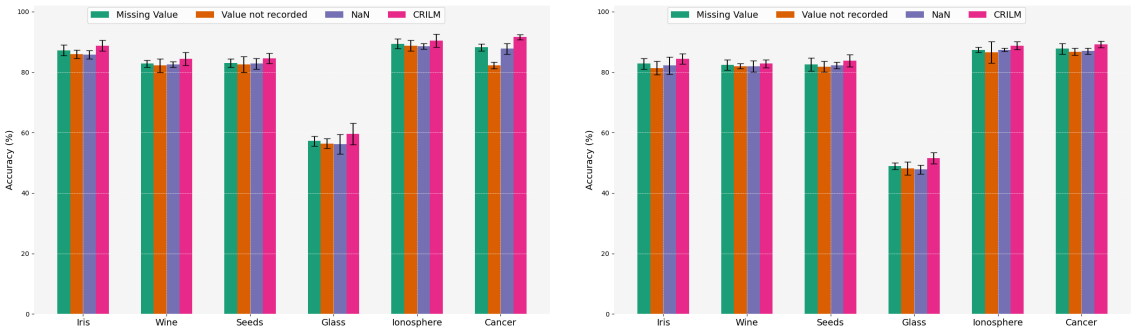
A further contribution of the missingness paper is that it makes the descriptor design problem concrete. CRILM does not merely replace every absent value with a universal placeholder. It generates feature-specific descriptors whose phrasing reflects the semantics of the underlying variable. Table 3.7 shows representative examples. The difference between *Malic acid quantity missing for this wine sample* and a generic token such as *NaN* is not simply stylistic. The first phrase tells the model what is absent and how it should be situated in the record. The second only signals absence at a syntactic level.

Table 3.7: Illustrative feature-specific missingness descriptors used in CRILM for selected datasets [1].

Dataset	Feature	Descriptor
Wine	Alcohol	Alcohol content not provided for this wine sample.
Wine	Malic acid	Malic acid quantity missing for this wine sample.
Wine	OD280/OD315	OD280/OD315 data missing for this wine sample.
Seeds	Area	Kernel area not provided.
Seeds	Asymmetry coefficient	Asymmetry coefficient information missing.
Iris	Petal width	Petal Width: Unavailable.

This feature-specificity was tested directly against generic alternatives. In the controlled MCAR experiments, CRILM compared feature-specific descriptors

with generic strings such as *NaN*, *Missing value*, and *Value not recorded*. The result was consistent across both LLaMA and FLAN-T5: feature-specific descriptors performed best, while generic placeholders degraded downstream accuracy [1]. Among the generic options, *Missing value* was usually the least harmful, but it still underperformed context-aware phrasing. This is a more precise version of the same lesson first observed in the educational ICMLA experiments. It shows that what matters is not merely using language instead of numbers, but using *the right language*.



(a) Descriptor comparison with LLaMA (b) Descriptor comparison with FLAN-T5

Figure 3.8: Feature-specific descriptors outperform generic placeholders when missingness is verbalized for downstream language-model learning [1].

3.9.8 Why the missingness research matters

The missingness-focused line of research matters in the dissertation for three reasons. First, it provides the clearest controlled demonstration that missing values can be represented semantically rather than only estimated numerically. Second, it shows that this idea is not restricted to one educational dataset or one architecture family. Third, it reveals a design pattern that recurs throughout the later chapters: when a foundation model is given an input representation aligned with its pre-trained inductive biases, downstream generalization improves without requiring

unrestricted end-to-end adaptation.

For the dissertation narrative, CRILM therefore plays a transitional role. It extends the earlier contextualization work from *observed* LE sequences to the more difficult case where parts of the sequence are absent. It also anticipates the three-tier forecasting framework, where imputation becomes only one stage in a broader semantic pipeline, and it prefigures ConText-LE, where the central object of design is the representation itself. Put differently, CRILM does not merely solve a preprocessing problem. It strengthens the dissertation's general argument that reliability in LE modeling begins with how evidence is represented before any predictor is trained.

3.9.9 The three-tier framework: imputation, selection, and forecasting

The three-tier missingness-aware forecasting framework expanded this idea into a fully structured pipeline [2]. Instead of treating forecasting as a standalone final-stage task, the framework decomposes qualitative LE modeling into three connected reasoning stages:

1. **Tier 1: LLM-informed imputation.** Missing responses are replaced with semantically meaningful descriptors, especially for patterns believed to be missing-not-at-random.

2. **Tier 2: Zero-shot feature selection.** LLMs assess the relevance of qualitative dimensions so that the final representation emphasizes informative non-cognitive signals rather than all available questions.

3. **Tier 3: Fine-tuned forecasting.** The selected and textualized trajectories are used for downstream binary engagement forecasting.

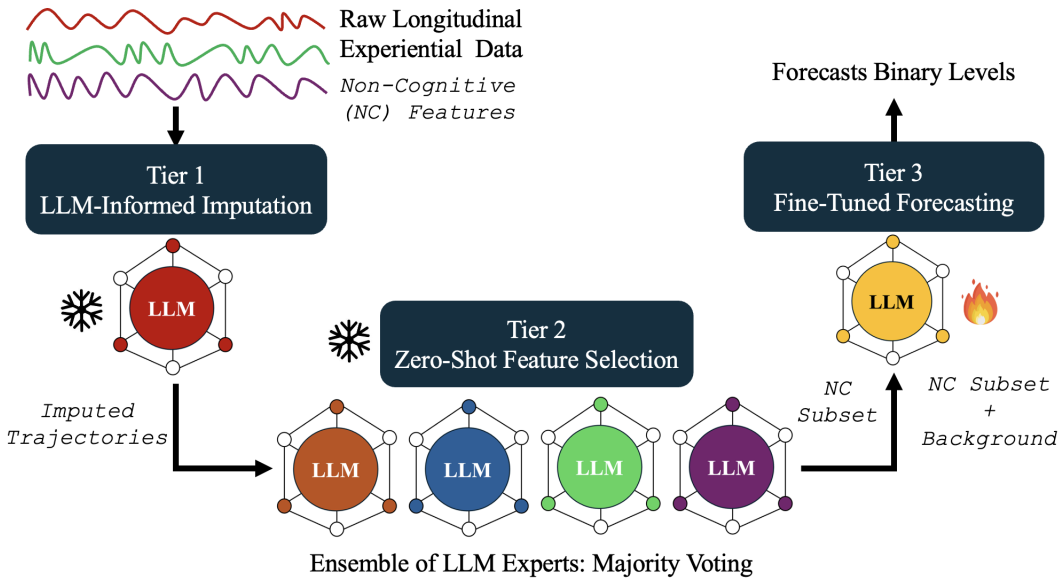


Figure 3.9: Three-tier framework for qualitative LE forecasting. The dissertation treats missingness handling, feature selection, and downstream classification as semantically connected stages rather than isolated preprocessing steps [2].

The significance of the three-tier design is broader than its individual modules. It shows that language models are not only useful as stronger classifiers over verbalized trajectories. They can also serve as semantic operators over the entire LE pipeline: interpreting missingness, curating relevant features, and then forecasting from the resulting representation. This is one of the clearest points in the dissertation where LLMs become a *pipeline technology* rather than a single model substituted for an older predictor.

3.9.10 Detailed empirical results for qualitative forecasting

The qualitative engagement experiments provide some of the strongest evidence in the dissertation for the claim that representation determines performance. Table 3.8 reports numeric baseline performance using only the selected non-cognitive subset features. Across the four engagement dimensions, the best

baseline balanced accuracy never rises above 55.5%, and macro-F1 stays in the narrow range from 39.5% to 54.0%. Even the most competitive classical sequence models struggle to extract stable signal from these sparse qualitative trajectories.

Table 3.8: Baseline performance across engagement dimensions using numeric non-cognitive subset features [2].

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

Table 3.9 then shows what changes when the same problem is recast into textual LLM inputs using only the non-cognitive subset features. The improvement is immediate and substantial. RoBERTa reaches 65.0% macro-F1 on LED, 70.5% on ASE, 69.5% on PSE, and 70.5% on AIVP. DistilBERT remains highly competitive, while decoder-only models exhibit more variable behavior across dimensions. Llama 7B, for instance, reaches 73.0% F1 on PSE but falls to 56.5% on LED, confirming again that generative capacity alone does not guarantee stable qualitative classification.

Table 3.9: LLM performance across engagement dimensions using selected non-cognitive features only [2].

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

The third configuration, shown in Table 3.10, appends background information to the selected non-cognitive subset. This is one of the clearest empirical

demonstrations in the chapter that contextualization is not decorative but structural. RoBERTa improves to 77.5% macro-F1 on LED, 73.0% on ASE, 73.5% on PSE, and 74.0% on AIVP. DistilBERT also improves substantially, and even decoder-only models become more stable when the qualitative trajectories are situated within background context. The mean macro-F1 rises from approximately 64.4% in the non-cognitive-only condition to roughly 69.5% once background context is added [2].

Table 3.10: LLM performance across engagement dimensions using selected non-cognitive features plus background information [2].

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

Ablation results deepen the interpretation. When all non-cognitive features are used instead of the zero-shot selected subset, performance declines across models, confirming that feature selection is not merely a convenience for reducing input length but a representation-level filter that removes noisy or weakly informative qualitative dimensions [2]. Likewise, when numeric baselines are trained on the same broad feature space, they still fail to match the best textual LLMs. The lesson is consistent with the larger dissertation: the gains do not come only from swapping in a larger model; they come from presenting the model with a better-curated, semantically coherent view of the trajectory.

3.9.11 Why this phase matters for the dissertation

This phase is one of the strongest validations of the dissertation's core premise. The performance gap between language-based and numeric approaches widens

precisely when the data becomes most qualitative, most missingness-rich, and most dependent on interpretation. The three-tier framework also makes visible a deeper architectural point that influences the rest of the dissertation: representation, curation, and prediction should not be thought of as strictly separate. In LE data, the choice of how to describe missingness, which features to foreground, and how to phrase the input already determines much of what the model can later learn. This is why the chapter treats missingness and feature selection not as support components, but as central to reliable representation.

3.10 Cross-Distribution Generalization and the ConText-LE Transition

By the time the dissertation reaches ConText-LE, the methodological question has shifted from stronger ID performance to robust behavior under distribution shift. This is the point where the LLM stage matures into a framework for generalization rather than a collection of promising task-specific improvements [3]. Earlier studies had shown that language models outperform numeric baselines on small, heterogeneous, behaviorally rich tasks. ConText-LE asks the more difficult question: do those gains survive when the model is trained on one temporal or contextual regime and then evaluated on another?

3.10.1 Cross-distribution generalization as the central reliability problem

In LE applications, distribution shift is the rule rather than the exception. Real deployment typically involves new semesters, new cohorts, new institutions, later time periods, or different participant groups. Formally, the forecasting problem is no longer confined to a single training distribution. Let $P_{\text{src}}(X, Y)$

Table 3.11: Selected empirical milestones across the language-model stage of the dissertation.

Study	Representative Result	Setting	Interpretation
Early forecasting with FLAN-T5	Up to 77% accuracy at 2 weeks and 89% at 8 weeks with full contextualization	Small-data educational performance forecasting	Pre-trained language models can make useful early predictions when trajectories are verbalized and contextualized [26].
Broader educational LE modeling	Up to 83.6% week-2 accuracy with tri-modal LLaMA input and 96.8% in the 4-week setting	Hybrid cognitive, non-cognitive, and background educational LE data	Performance improves when modalities are represented jointly and enriched with temporal cues and semantically meaningful missingness descriptors [20].
CRILM missingness-aware representation	Up to +10.0 points with LLaMA and +13.6 points with FLAN-T5 on challenging MNAR settings	Controlled MCAR/MAR/MNAR classification with semantically verbalized missing values	Missingness is often better treated as contextual evidence than as scalar repair, and language priors can improve downstream robustness across model families [1].
Lecture engagement forecasting	BERT-base improves from 66.88% to 77.07% accuracy when background features are added	Sparse qualitative engagement prediction	Context enrichment matters, and encoder-only models can be especially effective for qualitative classification [54].
Three-tier qualitative forecasting	RoBERTa reaches 77.5% macro-F1 with selected NC + background inputs	High-dimensional qualitative LE data with missingness and feature sparsity	LLMs are strongest when missingness is semantically handled, weak features are filtered, and the final representation remains contextually rich [2].
ConText-LE	Meta-Narrative + Prospective Narrative Generation reaches 67.40% OOD accuracy on GLOBEM, 67.19% on LifeSnaps, and 64.86% on MFAFY	Cross-distribution forecasting on three distinct LE benchmarks	Generalization depends on narrative representation and output formulation, not on model adaptation alone [3].

denote the source distribution and $P_{\text{tgt}}(X, Y)$ the target distribution. Then the cross-distribution problem can be written as

$$P_{\text{src}}(X, Y) \neq P_{\text{tgt}}(X, Y), \tag{3.26}$$

with the objective of learning a predictor

$$f_\theta : X \to Y \tag{3.27}$$

that remains reliable under both source and target conditions. In LE data, the shift may be cross-temporal, cross-cohort, cross-institutional, or cross-participant [3, 16, 28].

This matters because the failure of traditional approaches under shift is severe. The GLOBEM benchmark showed that nine previously published longitudinal behavior-modeling algorithms and multiple domain-generalization approaches remained near chance in OOD evaluation, with the best published baseline around 52.8% OOD accuracy [16]. That result became a reference point for the dissertation because it made clear that high ID performance says little about real transfer in behavioral forecasting.

3.10.2 Why text-only LLM forecasting still needed a new representational strategy

The earlier studies in this chapter already established that verbalization improves contextual reasoning. However, not all textualizations are equally effective under shift. A raw sequence dump, an aggregate summary, and a contextually synthesized narrative expose very different forms of evidence to the same base

model. ConText-LE therefore turns representation into the explicit variable of interest.

Let $X_{i,s:s+k-1}$ denote a k-week LE window. A textual representation operator $\mathcal{R}$ converts this structured window into a language input

$$X^{\text{text}}_{i,s:s+k-1} = \mathcal{R}(X_{i,s:s+k-1}). \tag{3.28}$$

The key insight is that $\mathcal{R}$ is not a neutral conversion step. It determines how much temporal structure is preserved, whether the model sees raw details or synthesized interpretations, and whether semantically meaningful relations are foregrounded or obscured. ConText-LE compares four such operators:

1. **Complete Sequence**, which directly verbalizes detailed temporal observations;

2. **Statistical Summary**, which collapses the window into aggregate feature statistics;

3. **Natural Language String**, which lists values over time in a more sentence-like form; and

4. **Meta-Narrative**, which synthesizes higher-level behavioral meaning across the entire window [3].

This shift in emphasis is essential for the dissertation. The research question is no longer merely whether an LLM helps. It is whether the model is given the right *kind* of textual evidence and the right output objective to learn abstractions that transfer across time and context.

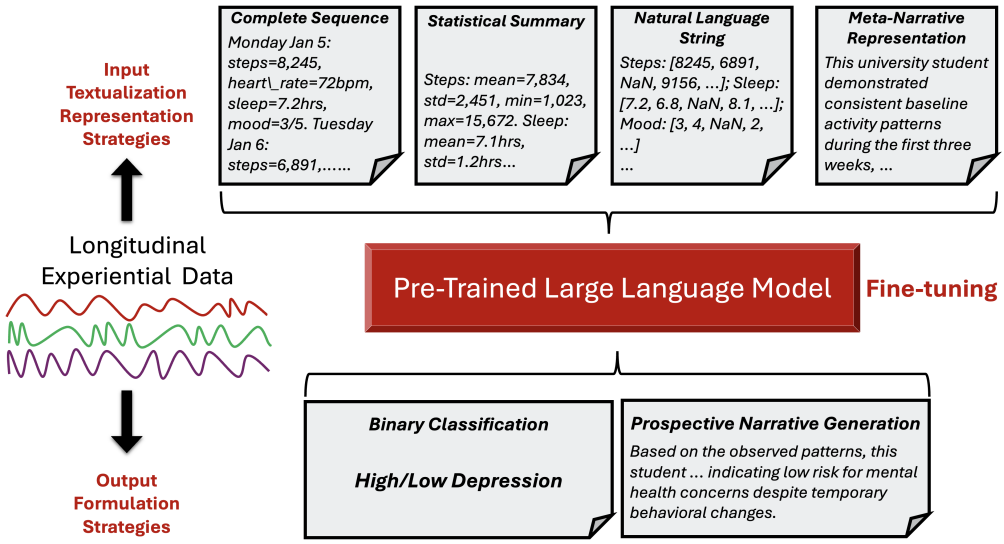


Figure 3.10: ConText-LE framework. The framework makes both input representation and output formulation explicit design choices, showing that OOD performance depends on how trajectories are textualized and what the model is asked to generate [3].

3.10.3 Meta-Narrative representation in detail

Meta-Narrative is the most conceptually important input representation developed in the chapter. It is motivated by the idea that LLMs reason more effectively when the input is not a raw textualized dump of values, but a semantically coherent account of what matters in the trajectory. Rather than preserving every detail at the same level of salience, Meta-Narrative foregrounds the most important patterns, temporal shifts, and cross-feature relations in a higher-level summary that still remains faithful to the observed evidence [3].

This representation can be understood as a two-stage synthesis process. First, feature-local patterns are analyzed:

$$p_{i,j} = \mathcal{A}_j \left(x_{i,s,j}, x_{i,s+1,j}, \ldots, x_{i,s+k-1,j} \right), \tag{3.29}$$

where $p_{i,j}$ summarizes the temporal behavior of feature j over the input window. Second, these local pattern summaries are integrated into a coherent narrative:

$$m_i = \mathcal{N}(p_{i,1}, p_{i,2}, \ldots, p_{i,F}, \mathcal{C}_i), \tag{3.30}$$

where $\mathcal{C}_i$ denotes contextual information and m_i is the final Meta-Narrative. In the actual implementation, GPT-4o is used to carry out this two-stage prompting process: feature pattern analysis first, contextual narrative synthesis second [3].

The value of this design is that it asks the model to reason over *behaviorally organized evidence* rather than over a lengthy sequence of weakly prioritized raw observations. The representation thus raises the semantic floor of the task. Instead of requiring the model to infer which local values matter and how they interact, part of that interpretive work is already performed before fine-tuning.

3.10.4 Output formulation as a generalization variable

ConText-LE also shows that representation alone is not enough. The output objective matters. Let the LLM be denoted by g_θ and the final output operator by $\mathcal{O}$. Then the overall prediction pipeline may be expressed as

$$\hat{y}_{i,s+k} = \mathcal{O}\big(g_\theta(X^{\text{text}}_{i,s:s+k-1})\big). \tag{3.31}$$

ConText-LE compares two choices for $\mathcal{O}$:

1. **Binary Classification**, where the model directly predicts a discrete label; and
2. **Prospective Narrative Generation**, where the model produces a future-oriented narrative from which a binary forecast can later be extracted [3].

This distinction is deeper than a change in output format. Binary classification compresses the supervision signal into a single label. Prospective Narrative Generation instead trains the model to produce a sequence

$$\hat{n}_i = g_\theta(m_i), \tag{3.32}$$

with token-level loss

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, m_i), \tag{3.33}$$

which better aligns with how LLMs are pretrained to operate. The intuition is that the model is encouraged to organize its internal reasoning around a coherent future-oriented account rather than a single hard label. This makes the task more natural for a generative language model and, as the results show, often more robust under distribution shift.

3.10.5 Main forward-direction results across GLOBEM, LifeSnaps, and MFAFY

The main forward-direction results from ConText-LE are given in Table 3.12. These are among the most important results in the entire LLM stage of the dissertation because they directly compare representation strategies and output formulations under OOD evaluation across three qualitatively different datasets.

Several observations are worth emphasizing in detail. On GLOBEM, the move from Binary Classification to Prospective Narrative Generation changes the performance regime entirely. Complete Sequence rises from 51.16% to 65.94% OOD accuracy, Statistical Summary from 51.11% to 62.43%, Natural Language String from 52.64% to 66.44%, and Meta-Narrative from 55.12% to 67.40%. This is not a

Table 3.12: Forward-direction cross-distribution generalization results ($T \rightarrow T'$) across all datasets. Bold indicates the best performance for each dataset and output formulation [3].

Dataset	Shift	Input Strategy	In-Distribution (ID)				Out-of-Distribution (OOD)			
			Acc	P	R	F1	Acc	P	R	F1
GLOBEM	Years 1&2 → Years 3&4	*Output: 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: 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: 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: 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: 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: 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**

marginal gain; it is a reformulation effect. Within the same generative setting, Meta-Narrative further surpasses the other representations, yielding 69.40% OOD F1 and establishing the strongest text-only result in the chapter on this benchmark [3].

On LifeSnaps, the representation effect is even more dramatic. Under Binary Classification, Statistical Summary achieves strong ID performance (86.96% F1) but collapses to 32.26% OOD F1, revealing severe over-specialization to the source regime. Meta-Narrative, by contrast, attains 56.67% OOD F1 even before the generative reformulation is applied. Once the task is shifted to Prospective Narrative Generation, Meta-Narrative reaches 67.19% OOD accuracy and 68.66% OOD F1, while Natural Language String also improves strongly to 60.00% F1. This pattern is important for the dissertation because it shows that cross-distribution robustness is not reducible to preserving more detail; it depends on *what kind of abstraction* the representation performs [3].

On MFAFY, the effect is subtler but still consistent. Under Binary Classification, Meta-Narrative yields 60.86% OOD accuracy and 64.96% F1, clearly ahead of the alternatives. Under Prospective Narrative Generation, it rises to 64.86% OOD accuracy while maintaining a competitive 64.14% F1. The more modest F1 gain relative to GLOBEM and LifeSnaps suggests that the educational engagement task is structurally different: some information that supports balanced classification is already well captured by the discriminative Meta-Narrative input, while the generative formulation primarily boosts transfer in terms of decision accuracy rather than uniformly across all metrics.

3.10.6 Bidirectional evaluation and reverse-direction results

A major strength of ConText-LE is that it does not stop at a single train-test direction. Whenever the data permits, the evaluation is performed in both

directions:

$$D_A \rightarrow D_B, \qquad D_B \rightarrow D_A, \tag{3.34}$$

where D_A and D_B denote two distinct temporal or contextual periods. This is important because it helps determine whether the improvement stems from genuine transferability or from direction-specific artifacts.

Table 3.13 summarizes the average and standard deviation of Meta-Narrative performance across both directions. The numbers are revealing. On GLOBEM, Binary Classification remains low and unstable relative to Prospective Narrative Generation. On LifeSnaps, Prospective Narrative Generation achieves both the highest average accuracy and the most stable F1. On MFAFY, the bidirectional means show that narrative generation is still advantageous in average accuracy, although the F1 differences become smaller and more sensitive to direction [3].

Table 3.13: Average (μ) and standard deviation (σ) of OOD performance across bidirectional experiments for Meta-Narrative input [3].

	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** ± 0.67	**67.92** ± 1.48	**69.31** ± 2.12	**68.94** ± 0.29	**67.67** ± 2.81	**64.07** ± 0.07

The complete reverse-direction tables are shown below because they are important for understanding how stable the representational advantages remain under a flipped temporal configuration. Their inclusion in the dissertation chapter matters: they make the generalization claim visibly harder to dismiss as a one-way temporal effect.

The reverse-direction tables reinforce several points that are easy to miss if only the forward direction is discussed. For GLOBEM, Meta-Narrative remains the strongest OOD representation under both formulations, but the largest jump again

Table 3.14: GLOBEM reverse-direction results ($T' \rightarrow T$: Years 3&4 $\rightarrow$ Years 1&2) [3].

	ID (Year 3&4 Test)				OOD (Year 1&2 Test)			
Input Strategy	Acc	P	R	F1	Acc	P	R	F1
Output: 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: 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 3.15: LifeSnaps reverse-direction results ($T' \rightarrow T$: Last 2 Months $\rightarrow$ First 2 Months) [3].

	ID (Last 2 Months Test)				OOD (First 2 Months Test)			
Input Strategy	Acc	P	R	F1	Acc	P	R	F1
Output: 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: 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**

comes from the switch to Prospective Narrative Generation: OOD accuracy rises from 55.08% to 68.75%. For LifeSnaps, the reverse direction produces the largest single absolute improvement observed in ConText-LE: Meta-Narrative improves from 56.36% to 71.43% OOD accuracy when the formulation changes from Binary Classification to narrative generation. For MFAFY, the asymmetry between directions becomes especially informative. Training on the shorter, more constrained Year-2 period yields stronger OOD transfer to Year 1 than the forward direction,

Table 3.16: MFAFY reverse-direction results ($T' \rightarrow T$: Year 2 $\rightarrow$ Year 1) [3].

	ID (Year 2 Test)				OOD (Year 1 Test)			
Input Strategy	Acc	P	R	F1	Acc	P	R	F1
Output: 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: 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**

with Meta-Narrative reaching 68.20% OOD accuracy under Binary Classification and 70.49% under narrative generation. These are not merely numerical extras; they show that the representation is learning something more stable than one-way temporal idiosyncrasies.

3.10.7 Comparison against non-LLM time-series baselines and LLM ablations

To situate ConText-LE relative to more conventional temporal models, the framework also compares against PatchTST and iTransformer. Table 3.17 shows that these baselines underperform consistently across all three datasets, particularly in OOD settings. On GLOBEM, PatchTST reaches only 49.16% OOD F1 and iTransformer 51.07%, far below the 69.40% OOD F1 achieved by Meta-Narrative plus Prospective Narrative Generation. Similar gaps appear on LifeSnaps and MFAFY [3].

The framework further examines whether the gains arise simply because of the chosen base model. Table 3.18 reports an architecture ablation on GLOBEM using the strongest ConText-LE configuration. Llama 3.1 8B Instruct remains the best overall model, but the gap to Mistral-7B-Instruct-v0.3 is not overwhelm-

Table 3.17: Comparison with non-LLM time-series baselines across ID and OOD settings [3].

Dataset	Model	ID		OOD	
		Acc	F1	Acc	F1
GLOBEM	PatchTST	53.58	53.01	49.88	49.16
	iTransformer	54.61	54.61	51.06	51.07
LifeSnaps	PatchTST	47.83	47.83	43.75	40.47
	iTransformer	52.17	48.83	48.44	47.99
MFAFY	PatchTST	67.39	64.34	50.57	44.31
	iTransformer	53.26	52.47	44.29	42.42

ing. Falcon-7B performs substantially worse, and the Llama 3.1 8B base model trails its instruction-tuned counterpart. This pattern suggests that instruction tuning matters and that model family differences are meaningful, but neither factor overwhelms the representational effect already established by the earlier comparisons.

Table 3.18: LLM architecture ablation on GLOBEM under the strongest ConText-LE setting [3].

LLM Architecture	ID		OOD		ID-OOD Gap
	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
Llama 3.1 8B Base	63.68	63.35	60.99	59.91	3.44

3.10.8 Why ConText-LE is the turning point of the LLM chapter

ConText-LE is the turning point at which the dissertation's LLM stage becomes explicitly a *generalization* stage. The earlier work established that language-based representation helps on small, difficult, heterogeneous tasks. ConText-LE demonstrates that the same representational logic also governs robustness under distribution shift. It is therefore not treated here as a standalone paper but as

the culmination of a line of reasoning: raw numerical forecasting gave way to verbalization; verbalization became contextualization; contextualization became semantically informed handling of missingness and feature relevance; and that, in turn, matured into narrative representation and future-oriented output design for cross-distribution transfer.

This is also the point where the text-only paradigm reaches its strongest form within the dissertation. Meta-Narrative plus Prospective Narrative Generation gives the best text-only OOD performance across GLOBEM, LifeSnaps, and MFAFY, materially improves over non-LLM temporal baselines, and does so with relatively small ID-OOD gaps compared with weaker textualizations [3]. At the same time, the chapter must also be honest about what remains unresolved: these gains are achieved while still reasoning over one-dimensional text. The next section makes that limitation explicit.

3.11 Synthesis: What the LLM Stage Contributes to the Dissertation

Across this chapter, the studies contribute several durable ideas to the overall dissertation.

First, they establish *verbalization as representational alignment*. LE data can be made more learnable by transforming it into language that better matches the inductive biases of foundation models. This is not a trivial engineering trick. It is a general principle about how structured human-centered data can interface with pre-trained models.

Second, they establish *contextualization as a forecasting principle*. The gains from distal and non-cognitive variables are not best understood as simple feature

additions. They reflect the fact that human trajectories are situated. Background and concurrent behavioral context can change the meaning of the same surface measurement.

Third, they broaden the dissertation's notion of what an LE pipeline is. By the time the research reaches qualitative engagement forecasting, LLMs are supporting descriptor-based imputation, feature selection, and prediction. The pipeline is therefore no longer a sequence of numeric preprocessing heuristics followed by a final model. It becomes a semantically mediated reasoning system.

Fourth, they show that *missingness is representational*. The wording of missing-value descriptors measurably changes downstream performance, and the three-tier framework demonstrates that semantically informed imputation is not merely a repair trick but part of a reliable input design.

Fifth, they demonstrate that *output design matters*. Prospective Narrative Generation is not simply a more verbose way to predict a label. It is a task formulation that aligns better with how language models encode and express contextual relationships, especially under shift.

Sixth, they show that *generalization is representational*. The strongest OOD improvements in ConText-LE come from better narrative abstraction and better-matched output formulation, not only from replacing one base model with another. This directly supports the central claim of the dissertation.

3.12 The Limits of Text-Only LLM Modeling

A dissertation chapter should not end at the point where a method works better than what came before. It should identify what remains unresolved. Despite the major advances described above, text-only LLM modeling still has limitations

that prevent it from being the final answer to LE generalization.

3.12.1 Serialization distorts structured temporal data

The strongest limitation is structural. LE data often has an underlying two-dimensional organization: features by time. When that structure is converted into a one-dimensional token sequence, local relationships can become distant in the serialized representation. If there are F features observed across T time steps, then a straightforward serialization has length roughly

$$L_{\text{text}} \propto F \times T, \tag{3.35}$$

and, more importantly, two adjacent time points of the same feature can become separated by all intervening tokens from the other features. In a feature-major serialization, the gap between neighboring temporal observations may satisfy

$$\text{dist}_{\text{text}}\big((f,t),(f,t+1)\big) \geq F - 1. \tag{3.36}$$

This captures the core representational distortion of text-only modeling: temporal adjacency in the original structure does not imply positional adjacency in the token sequence.

3.12.2 Narrative abstraction still compresses away some structure

Meta-Narratives improve semantic organization, but they necessarily compress the original trajectory. That compression is useful for generalization because it suppresses some source-specific noise. At the same time, it can remove weak but predictive details, fine-grained cross-feature synchronization, or local structural

cues that become important once the data is more richly multimodal. In other words, the strongest text-only representation is still a summary of a richer object.

3.12.3 Pipeline dependence on external generation steps

The strongest text-only methods in this chapter rely on high-quality generation steps for textualization, target narrative construction, and in some cases label extraction. Methodologically, this is acceptable and often powerful. Scientifically, however, it introduces cost, latency, and pipeline dependency. These issues do not invalidate the approach, but they motivate a search for representations that preserve more structure without requiring every relationship to be rewritten into long text.

3.12.4 Why the next chapter must move to vision-language modeling (VLM)

These limitations motivate the dissertation's next transition. If the problem with text-only modeling is that it linearizes a richer Feature$\times$Time object, then a natural next question is whether some of that lost structure can be preserved through a second modality. This is exactly the move undertaken in the next chapter: visual encodings of LE trajectories are introduced so that temporal locality, cross-feature co-occurrence, and structural regularities can be preserved spatially rather than only linguistically [21, 22]. The move to vision-language modeling (VLM) is therefore not an abandonment of the LLM stage. It is the next logical response to the structural ceiling reached by the strongest text-only methods.

3.13 Chapter Summary and Bridge to Chapter 4

This chapter showed that the dissertation's move to LLMs was not simply a move to larger models. It was a move to a different representational paradigm for LE data. The chapter traced that progression across early forecasting, contextualized educational modeling, broader LE representation, semantically informed treatment of missingness, qualitative engagement forecasting, and finally ConText-LE. Across those stages, the dissertation established that language models become most useful when trajectories are verbalized in ways that preserve context, foreground salient behavioral patterns, and align the task with the model's pretrained capacities.

These findings primarily address **RQ2** and **RQ3**, and they support **H1**, **H2**, and **H3**: contextualized verbalization improves LE forecasting over conventional numeric baselines, semantically informed handling of missingness improves downstream modeling, and narrative output design strengthens OOD generalization.

At the same time, the chapter also identified a limit. Even the best text-only LE modeling pipeline ultimately reasons over linearized sequences. Missingness can be described more faithfully, context can be integrated more coherently, and future states can be formulated as narratives rather than labels; yet the underlying Feature$\times$Time structure is still compressed into one dimension. This leaves open a crucial question: can the structural relations that text weakens be preserved through another modality without sacrificing the semantic strengths gained in the LLM stage?

The next chapter answers that question by moving to vision-language modeling (VLM). It examines how visual encodings of LE trajectories can complement language-based meta-narratives so that structure is preserved spatially while inter-

pretation remains semantically grounded. In the dissertation arc, this is the point where the dissertation moves from the strongest form of text-only generalization to the first genuinely multimodal answer to reliable LE modeling under distribution shift.

Chapter 4

Vision-Language Modeling (VLM) of LE Data

4.1 Introduction

The previous chapter established that LLMs materially improve LE forecasting when trajectories are represented through semantically grounded narratives rather than flattened numeric inputs. In particular, ConText-LE showed that *Meta-Narrative* representations and *Prospective Narrative Generation* can substantially improve cross-distribution performance by aligning forecasting with the contextual reasoning strengths of language models [3]. That result marked the strongest purely language-based stage of this dissertation. Yet it also exposed a boundary that becomes central here: even the best text-only LE representation remains a *one-dimensional token sequence* derived from data that is natively organized across both features and time.

This chapter takes the next step in the dissertation arc by asking a precise representational question: *what information is still being lost when LE trajectories are modeled through text alone, even if that text is semantically rich?* The answer developed in this chapter is that text-only modeling still weakens temporal locality, cross-feature grouping, and the visual geometry of change. Narrative abstraction helps the model understand *what* the trajectory means, but it does not fully preserve

how that trajectory is organized. In LE data, that organization often matters. A week-to-week decline, synchronized movement across multiple features, a burst of volatility, or the contrast between stability and sudden rupture are not merely numerical details. They are structured temporal patterns whose interpretation depends on adjacency, shape, and co-occurrence. When those patterns are forced into a long token stream, part of the relational signal is inevitably compressed or dispersed [21, 67, 68].

This observation motivates the transition from language-only modeling to *vision-language modeling (VLM)*. The goal is not to abandon language. Language remains essential because it provides semantic abstraction, contextual interpretation, and the narrative framing that proved so effective in the previous chapter. Instead, this chapter argues that the strongest LE representation at this stage is *multimodal*: text for interpreted behavioral meaning, and visual encoding for preserving the two-dimensional Feature $\times$ Time structure that text-only serialization weakens. The central framework, LE-Viz, operationalizes this idea by pairing textual Meta-Narratives with structured visual encodings such as per-feature line charts and heatmap chains, and then learning to forecast future states through prospective narrative generation within a vision-language model (VLM) [21].

The chapter makes four major arguments. First, it shows why text-only modeling remains structurally incomplete even after the semantic advances of narrative-based LLM representations. Second, it develops the conceptual and methodological basis for treating visual encoding as a structure-preserving representation of LE data rather than a mere display format. Third, it presents LE-Viz in detail, including its problem formulation, multimodal input construction, visual design choices, training setup, and empirical evaluation across three cross-distribution forecasting benchmarks. Fourth, it analyzes the resulting evi-

dence to clarify when and why visual encoding helps, why not all VLMs benefit equally, and why multimodal representation is necessary but still not the final answer of the dissertation.

The emphasis of this chapter is therefore deliberately different from the previous one. The previous chapter was primarily about semantic contextualization, missingness-aware language modeling, and narrative abstraction. This chapter is about *structure preservation*. It shows that once LE forecasting is reframed as a multimodal representation problem, visual encoding becomes more than an auxiliary modality. It becomes a corrective to the representational collapse created by one-dimensional serialization. At the same time, the chapter remains careful not to overstate the conclusion. The results show that multimodality helps substantially, but they also reveal important limitations related to architectural alignment, data type, computational cost, and modality imbalance. Those limits are kept in view throughout, because the purpose of this chapter is not only to present LE-Viz in detail, but also to prepare the ground for the more constrained multimodal framework that follows in the next chapter.

4.2 From Narrative Generalization to Multimodal Representation

The transition from ConText-LE to LE-Viz is best understood not as a change of model family alone, but as the consequence of a deeper representational diagnosis. The previous chapter demonstrated that better semantic abstraction and better output formulation can materially improve out-of-distribution (OOD) forecasting. Meta-Narrative input outperformed simpler textualizations because it synthesized temporal patterns and contextual interactions into a form that LLMs could interpret more effectively. Prospective Narrative Generation improved over direct

binary classification because it aligned supervision with the generative prior of the model and provided denser token-level learning signals [3]. Those contributions established a strong language-centered solution to LE forecasting.

However, the same chapter also made clear that even strong narrative representations still rely on a sequential interface to inherently structured data. LE trajectories are not born as token streams. They typically begin as matrices or collections of temporally aligned heterogeneous signals. In mental health forecasting, for example, daily or weekly measures of activity, sleep, phone use, mobility, and self-reported state coexist across a shared timeline. In educational LE data, qualitative and quantitative indicators of engagement, confidence, identity, and behavior evolve jointly over time. What the model must learn is therefore not merely the meaning of isolated values, but the structure of their co-development. This includes within-feature trajectories, between-feature co-variation, and the distinction between persistent trends and transient perturbations. Narrative abstraction can summarize such relations, but it cannot preserve every structural property equally well.

The core limitation can be stated simply. When a Feature $\times$ Time representation is serialized into text, one of two things happens. Either the serialization favors feature continuity, in which case time-local co-occurrences across features are weakened, or it favors time-step grouping, in which case long-range trajectories of individual features are broken apart. In both cases, the model must reconstruct two-dimensional relationships from a linear token order that was imposed only for compatibility with text processing. This is a workable compromise for some tasks, but it becomes increasingly fragile as trajectories become longer, more heterogeneous, and more dependent on pattern shape rather than isolated value identity [21].

The move to multimodal LE modeling therefore begins with a sharper dissertation claim:

> *Semantic contextualization is necessary for LE forecasting, but semantic contextualization alone is not sufficient when the native structure of the data is itself predictive.*

Visual encoding becomes attractive at exactly this point because it provides a way to preserve structural relationships through spatial organization while allowing language to remain responsible for interpretation. LE-Viz therefore does not replace the earlier narrative approach. It *builds on it*. The textual Meta-Narrative from ConText-LE is retained as the semantic channel. What changes is that the same trajectory is also encoded visually so that the VLM can access structural cues that text alone obscures. This design choice is conceptually important for the dissertation as a whole: it marks the point where the dissertation moves from better textualization to *complementary multimodal representation*.

4.3 Why Text-Only Modeling Remains Structurally Incomplete

To formalize the representational issue, let an LE trajectory for individual i be represented as

$$X_i \in \mathbb{R}^{F \times T}, \tag{4.1}$$

where F denotes the number of features and T denotes the number of time steps within the observation window. Although this matrix notation is an abstraction over heterogeneous inputs, it captures the key idea that LE data preserves at least two kinds of locality:

1. **Temporal locality:** adjacent values within the same feature form short-range local patterns such as gradual drift, oscillation, discontinuity, or stability.

2. **Cross-feature locality:** values observed at the same time step can form behaviorally meaningful configurations across features.

In a purely textual pipeline, the model does not directly receive X_i. It instead receives a serialization

$$s_i = \mathcal{S}(X_i), \tag{4.2}$$

where $\mathcal{S}(\cdot)$ converts the native structure into a token sequence. The semantic gains of the previous chapter come from choosing a stronger $\mathcal{S}(\cdot)$, such as Meta-Narrative rather than complete raw verbalization. But even the strongest $\mathcal{S}(\cdot)$ remains a mapping from a two-dimensional organization into a one-dimensional order.

This mismatch becomes clearer when the trajectory contains many features observed across multiple time steps. If the sequence is serialized feature by feature, then values that were synchronized in time become separated by many intervening tokens. If it is serialized time step by time step, then the long-range continuity of each individual feature becomes harder to maintain. A simplified expression for the resulting token burden is

$$L_{\text{text}} \propto F \times T, \tag{4.3}$$

which means that increasing either feature dimensionality or temporal horizon lengthens the sequence and weakens the correspondence between token proximity and native structural proximity.

LE-Viz sharpens this intuition further. When LE data with F features across T time steps is verbalized in a detailed sequential form, temporally adjacent values

of the same feature may be separated by at least $F - 1$ intervening tokens. Even with self-attention, this forces the model to recover local temporal structure by searching across token positions whose ordering was not designed to preserve the structure in the first place. The problem is therefore not merely that the sequence is long. It is that the *geometry of the signal* no longer aligns well with the geometry of the representation [21].

This structural incompleteness matters because many LE forecasting tasks are driven by pattern form rather than isolated values. Consider several recurring cases:

- a monotonic decline in sleep paired with growing variability in mood,
- synchronized improvement across activity and social interaction,
- sharp divergence between self-reported confidence and observed performance,
- or repeated alternation between stable periods and abrupt shocks.

Such relationships can be described in text, and Meta-Narratives do exactly that at a high level. But those descriptions are already one step removed from the native arrangement of the data. They preserve interpretation, not full structure. The more the task depends on local shape, relative positioning, and joint temporal evolution, the more expensive it becomes to encode all of that faithfully through language alone.

This does not negate the LLM results from the previous chapter. On the contrary, it explains both their strength and their limit. Narrative-based LLMs succeed because they align better with contextual interpretation than traditional numeric models do. They remain limited because they still inherit the constraint

of text-only sequence processing. Visual encoding becomes attractive exactly at this boundary.

4.4 Why Visual Encoding Becomes Attractive

The justification for visual encoding in LE data rests on more than the practical availability of VLMs. It is grounded in a long tradition of work on visuospatial reasoning, diagrammatic representation, and multimodal learning. Visualizations do not merely present the same information in a different cosmetic form. They reorganize information so that some relationships become perceptually immediate through spatial proximity, alignment, grouping, slope, curvature, color continuity, and geometric contrast [67, 68]. In other words, visual form can change which relations are easy to detect.

This is particularly relevant for LE forecasting because many predictive regularities are inherently geometric. Trend direction can be perceived as slope. Volatility can be perceived as jaggedness or rapid color change. Persistence can be perceived as flatness or long coherent segments. Cross-feature synchronization can be perceived as aligned peaks, troughs, or simultaneous intensity shifts. These are not arbitrary visual metaphors. They are direct representations of temporal structure.

Suppose a visual encoding function maps the native LE trajectory into an image-like representation:

$$v_i = \mathcal{E}_{\text{vis}}(X_i). \tag{4.4}$$

If $\mathcal{E}_{\text{vis}}(\cdot)$ is designed so that temporal adjacency in X_i becomes spatial adjacency in v_i, then a vision encoder can process temporal relationships through local patch interactions rather than reconstructing them from scattered text positions. This is

the central representational advantage of the visual channel in LE-Viz.

The attraction of visual encoding is therefore strongest when three conditions hold:

1. the task depends on temporal shape or multi-feature interaction,
2. the underlying data has a meaningful Feature $\times$ Time organization,
3. and the model benefits from seeing local spatial structure rather than only receiving abstract verbal summaries.

The datasets used in this chapter satisfy these conditions to varying degrees. GLOBEM and LifeSnaps contain continuous sensor-driven trajectories whose progression is naturally visualizable. MFAFY is more qualitative and categorical, making it a useful boundary case: if visual encoding helps less there, that would reveal something important about when the visual channel contributes most.

At the same time, visual encoding is not introduced as a replacement for language. Text and image contribute different strengths. The textual channel remains better suited for contextual interpretation and semantic synthesis. The visual channel remains better suited for structure preservation and local pattern accessibility. The chapter therefore treats visual encoding as a *complement*, not a substitute.

4.5 LE-Viz: Vision-Language Modeling for LE Data

4.5.1 Problem formulation

LE-Viz addresses cross-distribution generalization for heterogeneous sequential data. Models are trained on a source distribution T and evaluated on a target distribution T' drawn from different cohorts, institutions, years, or time

periods [21]. For data collected from N individuals over K weeks, the dissertation follows a sliding-window formulation similar to the benchmark setup used in earlier chapters. For each individual i, a k-week observation window beginning at week s and ending at week $s + k - 1$ is used to forecast the subsequent week $s + k$.

In LE-Viz, the raw sequence is transformed into a multimodal pair

$$\left(X^{\text{text}}_{i,s:s+k-1},\ X^{\text{visual}}_{i,s:s+k-1}\right), \tag{4.5}$$

where the textual component provides semantic context and the visual component preserves structural organization. Rather than predicting a label directly, the model generates a prospective narrative for the future state:

$$f_\theta : \left(X^{\text{text}}_{i,s:s+k-1}, X^{\text{visual}}_{i,s:s+k-1}\right) \rightarrow y^{\text{text}}_{i,s+k}. \tag{4.6}$$

This preserves continuity with ConText-LE while expanding the input space beyond text.

4.5.2 Multimodal input transformation

The first stage of LE-Viz converts raw heterogeneous trajectories into complementary textual and visual forms. Conceptually, this stage performs two different kinds of abstraction over the same underlying data:

$$m_i = \mathcal{E}_{\text{text}}(X_i), \tag{4.7}$$

$$v_i = \mathcal{E}_{\text{vis}}(X_i), \tag{4.8}$$

where m_i is a high-level narrative representation and v_i is a structure-preserving visual representation.

Textual Meta-Narrative channel. LE-Viz retains the Meta-Narrative representation developed in ConText-LE as its textual component. This is important for two reasons. First, it provides semantic continuity across chapters, making LE-Viz a direct extension of the strongest prior text-only baseline rather than an unrelated multimodal system. Second, it prevents the visual channel from carrying the entire burden of interpretation. The Meta-Narrative distills global trends, feature interactions, contextual anomalies, and high-level behavioral implications into a compact narrative summary. In the LE-Viz implementation, GPT-4o is used to generate these summaries, and the same Meta-Narrative also serves as the strongest text-only baseline in the experiments [21].

Visual channel. The visual channel explicitly preserves feature dynamics through spatial organization. LE-Viz implements two visualization types:

- **Feature trajectory line charts**, which map time to the horizontal axis and feature value to the vertical axis, preserving slope, curvature, stability, and abrupt change.
- **Feature heatmap chains**, which map consecutive time steps to spatially ordered markers whose color intensity encodes value progression.

The two encodings highlight different visual affordances. Line charts make exact trend geometry particularly salient. Heatmap chains emphasize relative temporal progression through color continuity and intensity change. Both are designed so that nearby time steps appear in neighboring spatial regions, allowing a vision encoder to process temporal patterns through local receptive structure rather than through long-range token reconstruction.

Interleaved multimodal organization. A distinctive design choice in LE-Viz is the *interleaved* input format. Instead of presenting all text together and all images together, the framework places the global Meta-Narrative first and then organizes the remainder as feature-specific modules in which each brief textual summary is paired locally with its corresponding chart or heatmap. Abstractly, if $m_{i,j}$ and $v_{i,j}$ denote the feature-specific text and visual encoding for feature j, then the multimodal input can be written as

$$z_i = \left[m_i^{\text{global}}, (m_{i,1}, v_{i,1}), (m_{i,2}, v_{i,2}), \ldots, (m_{i,F}, v_{i,F})\right]. \tag{4.9}$$

This arrangement is not a cosmetic formatting decision. It is a representational hypothesis: local text-vision pairing should support more reliable cross-modal grounding than a layout that groups all visuals into a single combined image separated from their textual counterparts.

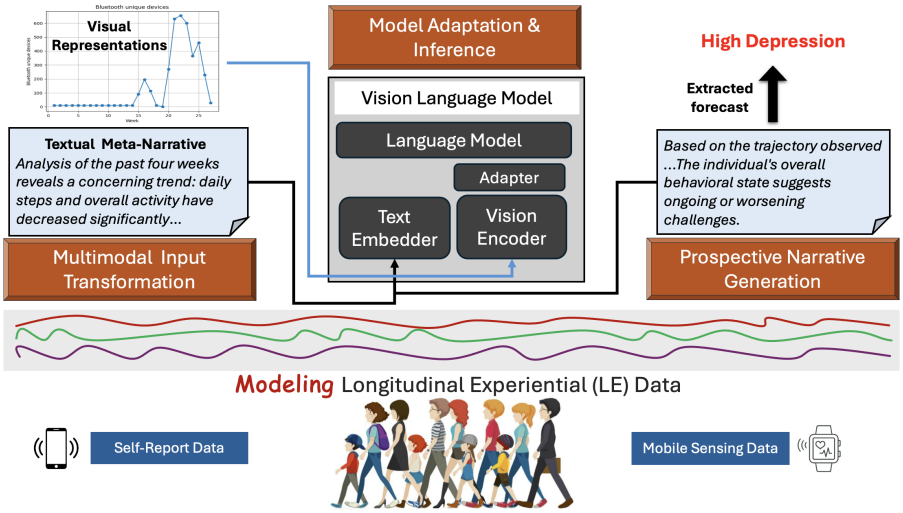


Figure 4.1: Overview of the LE-Viz framework. Raw heterogeneous LE trajectories are transformed into complementary textual Meta-Narratives and visual encodings, then processed by a VLM that generates a prospective narrative for the future state. This figure is reproduced from the LE-Viz source materials used to build this chapter [21].

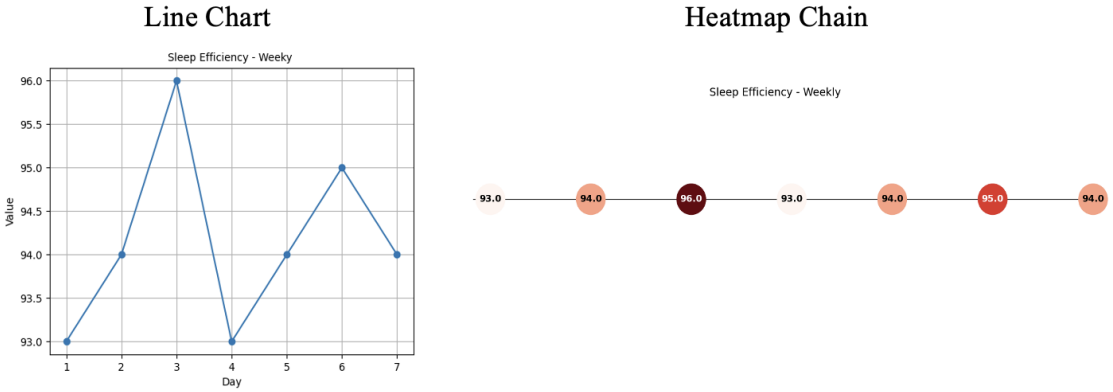


Figure 4.2: Examples of per-feature visual encodings used in LE-Viz. The left panel shows a feature trajectory line chart; the right panel shows a feature heatmap chain. The key design principle is that temporal progression is preserved as spatial progression [21].

4.5.3 Generative forecasting with a VLM

LE-Viz keeps the generative forecasting philosophy introduced in ConText-LE. Given the multimodal input, the VLM generates a prospective narrative

$$\hat{n}_i = g_\theta(v_i, m_i), \tag{4.10}$$

where $\hat{n}_i$ is a natural-language description of the likely future state. Training uses a causal language modeling objective:

$$\mathcal{L}_{\text{vlm}} = -\sum_{t=1}^{T_n} \log p_\theta(n_t \mid n_{<t}, v_i, m_i), \tag{4.11}$$

with T_n denoting target narrative length.

This choice matters for the dissertation because it isolates the representational question more cleanly. If LE-Viz had switched both the input representation and the output formulation, it would be harder to tell whether the gains came from multimodality or from a task reformulation. By keeping the prospective narrative

objective from the previous chapter, LE-Viz changes primarily the *input side*: it asks whether adding structure-preserving visual context on top of Meta-Narrative text improves cross-distribution forecasting.

4.5.4 Model adaptation and implementation choices

The main LE-Viz model uses LLaVA-NeXT 7B as the base VLM, with a SigLIP vision encoder and an instruction-tuned language backbone [69]. Fine-tuning is performed with LoRA, using a parameter-efficient adaptation strategy rather than fully updating the backbone. The reported configuration uses LoRA rank 64, alpha 128, AdamW optimization, a learning rate of 10^{-4}, a batch size of 4, 20 training epochs, and bfloat16 mixed precision. The text-only LLM baselines use Llama 3.1 8B Instruct with a lighter LoRA setup. Training was conducted on 8 NVIDIA A40 GPUs (48GB each) [21].

These details are important for interpretation. LE-Viz already represents a substantially heavier stage than the previous text-only chapter. The chapter therefore treats computational cost as part of the result, not just an implementation footnote. The multimodal gains are real, but they are achieved within a more expensive adaptation regime, which later becomes relevant when considering what kind of multimodal system is most sustainable and reliable for LE data.

4.6 Datasets, evaluation protocol, and baselines

LE-Viz is evaluated on three cross-distribution benchmarks chosen to test multimodal LE modeling across distinct domains and input types. Table 4.1 summarizes the settings.

The evaluation protocol follows the dissertation's general emphasis on source-

Table 4.1: Datasets used in the LE-Viz chapter. The table highlights why these benchmarks are useful together: GLOBEM and LifeSnaps emphasize continuous sensing-driven temporal dynamics, whereas MFAFY provides a harder qualitative educational LE setting with weaker native continuity in the visual channel [16, 20, 21, 28].

Dataset	Domain	Input characteristics	Cross-distribution split	Window / target	Approx. sequences
GLOBEM	Mental health forecasting	15 mobile sensing features including location, phone use, bluetooth, calls, physical activity, and sleep	Train on Years 1–2 from Institution A; evaluate OOD on Years 3–4 from Institution B	4-week window predicting subsequent week	$\sim$ 2226 train, $\sim$ 2023 OOD
LifeSnaps	Stress / anxiety forecasting	Multi-source wearable and EMA data with over 35 data types; subset used for binary anxiety forecasting	Train on first 2 months from one participant set; evaluate OOD on last 2 months from disjoint participants	1-week window predicting subsequent week	$\sim$ 112 train, $\sim$ 64 OOD
MFAFY	Educational engagement forecasting	Qualitative non-cognitive student trajectories with relevant engagement-related dimensions	Train on Year 1; evaluate OOD on Year 2	4-week window predicting subsequent week	$\sim$ 610 train, $\sim$ 350 OOD

to-target transfer. For each dataset, 15% of the data from the training period is reserved as an in-distribution (ID) test set, while the entire distinct target period is used as the OOD test set. This allows the chapter to compare not only overall performance, but also the size of the ID-to-OOD drop under each representational strategy [21].

The baselines are intentionally broad. They include:

- **Time-series models:** PatchTST and iTransformer, representing strong modern sequence models that operate on numerical time-series structure without language or vision priors [70, 71].
- **Text-only LLM baselines:** LLM-Complete, LLM-Summary, LLM-Clustered, and LLM-Meta, which vary the textualization strategy while using the same general generative paradigm established earlier in the dissertation [3, 20, 27,

72, 73].

- **Generic multimodal temporal baseline:** Time-VLM, included to test whether a generic vision-language time-series model transfers well to heterogeneous LE data without the purpose-designed representational choices of LE-Viz [74].

This baseline design is important. It means that LE-Viz is not being compared only against older numeric methods. It is also compared against the strongest text-only representation from the previous chapter and against a generic multimodal time-series architecture. The empirical question is therefore specific: *does purpose-designed visual encoding for LE data improve generalization beyond both text-only narrative modeling and generic multimodal forecasting?*

4.7 Main results across datasets

Table 4.2 presents the primary cross-distribution results. It is the central empirical table of this chapter because it places classical time-series models, text-only LLMs, a generic VLM baseline, and the two LE-Viz variants into a single comparison.

4.7.1 Results on GLOBEM

GLOBEM is the clearest demonstration of why the visual channel matters. The time-series baselines remain near chance under OOD evaluation, with PatchTST at 49.88% OOD accuracy and iTransformer at 51.06%. Time-VLM, despite using a vision-language architecture, performs similarly at 51.43% OOD accuracy and 50.04% F1. This is a revealing failure: adding multimodal machinery in a generic way does not by itself solve the LE representation problem.

Table 4.2: Cross-distribution generalization results for LE-Viz across all datasets. Bold indicates the best result within each dataset and evaluation regime. This table is adapted from the LE-Viz experiments and retained here because it is the core empirical evidence for the chapter [21].

Dataset	Shift	Method	In-Distribution (ID) Test				Out-of-Distribution (OOD) Test			
			Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
GLOBEM	Years 1&2 → Years 3&4	***Time-Series Models***								
		PatchTST	53.58	53.73	53.58	53.01	49.88	49.24	49.88	49.16
		iTransformer	54.61	54.61	54.61	54.61	51.06	51.07	51.06	51.07
		LLM Models								
		LLM-Complete	69.96	71.56	68.42	69.96	65.94	67.95	68.52	68.23
		LLM-Summary	69.51	72.22	67.24	69.65	62.43	65.97	63.57	64.75
		LLM-Clustered	70.05	71.30	69.37	70.32	66.44	67.92	69.09	68.50
		LLM-Meta	73.99	75.93	71.93	73.87	67.40	68.81	70.00	69.40
		VLM Models								
		Time-VLM	57.34	55.92	56.48	56.20	51.43	49.87	50.21	50.04
		LE-Viz (Meta + Heatmap)	76.68	72.83	71.28	72.04	70.98	73.19	73.13	73.16
		LE-Viz (Meta + Chart)	**79.37**	**83.33**	**80.77**	**82.03**	**72.86**	**75.58**	**74.28**	**74.92**
LifeSnaps	First 2 Months → Last 2 Months	***Time-Series Models***								
		PatchTST	47.83	47.83	47.83	47.83	43.75	41.23	43.75	40.47
		iTransformer	52.17	46.81	52.17	48.83	48.44	48.14	48.44	47.99
		LLM Models								
		LLM-Complete	58.82	77.78	58.33	66.67	54.84	50.00	57.14	53.33
		LLM-Summary	47.06	40.00	57.14	47.06	46.88	36.67	42.31	39.29
		LLM-Clustered	70.59	80.00	72.72	76.19	62.50	52.94	69.23	60.00
		LLM-Meta	64.71	77.78	63.64	70.00	67.19	63.89	74.19	68.66
		VLM Models								
		Time-VLM	49.28	47.11	47.86	47.48	47.36	45.02	45.71	45.36
		LE-Viz (Meta + Heatmap)	76.47	83.33	62.50	71.43	69.23	72.41	63.64	67.74
		LE-Viz (Meta + Chart)	**82.35**	**85.71**	**75.00**	**80.00**	**71.88**	**83.33**	**71.43**	**76.92**
MFAFY	Year 1 → Year 2	***Time-Series Models***								
		PatchTST	67.39	63.66	67.39	64.34	50.57	49.68	50.57	44.31
		iTransformer	53.26	51.77	53.26	52.47	44.29	43.05	44.29	42.42
		LLM Models								
		LLM-Complete	60.66	56.67	60.71	58.62	57.14	50.55	60.53	55.09
		LLM-Summary	57.38	48.28	56.00	51.85	53.43	52.02	52.94	52.48
		LLM-Clustered	63.93	62.96	69.37	60.71	62.86	57.47	64.10	60.61
		LLM-Meta	70.49	65.22	60.00	62.50	64.86	61.11	**67.48**	64.14
		VLM Models								
		Time-VLM	53.41	51.76	52.33	52.04	48.27	46.14	46.82	46.48
		LE-Viz (Meta + Heatmap)	72.13	69.57	61.54	65.31	65.43	62.94	56.96	59.80
		LE-Viz (Meta + Chart)	**77.05**	**75.76**	**80.65**	**78.12**	**66.57**	**64.47**	60.87	**62.62**

The text-only baselines perform considerably better. LLM-Complete, LLM-Summary, and LLM-Clustered all improve on the numerical models, while LLM-Meta from ConText-LE reaches the strongest text-only performance with 67.40% OOD accuracy and 69.40% F1. This confirms the central conclusion of the previous chapter: narrative abstraction already produces a major gain.

LE-Viz then pushes beyond that ceiling. The heatmap variant reaches 70.98% OOD accuracy and 73.16% F1. The chart variant performs best overall, achieving 72.86% OOD accuracy and 74.92% F1. Relative to the best text-only baseline, this is a gain of 5.46 percentage points in OOD accuracy and 5.52 points in OOD F1. Relative to the generic Time-VLM baseline, the gap is even larger. In practical terms, the result shows that the strongest gains come not from using a VLM generically, but from presenting LE data to the VLM in a representation specifically designed to preserve temporal and cross-feature structure.

The GLOBEM result is especially important because the benchmark was created precisely to expose cross-distribution brittleness in human behavior modeling. Achieving 72.86% OOD accuracy on this task means that visual-text multimodal representation is not just improving cosmetic performance on an easy dataset; it is addressing a benchmark whose purpose is to make generalization genuinely difficult [16, 21].

4.7.2 Results on LifeSnaps

LifeSnaps provides a second test in a different sensing-driven domain. Here too, the time-series baselines and Time-VLM remain weak under OOD evaluation, clustered in the mid-40% range. The strongest text-only baseline, again LLM-Meta, reaches 67.19% OOD accuracy and 68.66% F1.

LE-Viz maintains its advantage. The heatmap version achieves 69.23% OOD accuracy, while the chart version reaches 71.88% OOD accuracy and 76.92% F1. The improvement over LLM-Meta is therefore about 4.69 percentage points in OOD accuracy and more than 8 points in OOD F1. This is a particularly instructive result because it shows that the visual channel is not merely repeating the information already encoded in the narrative. If it were, the multimodal model would tend to

match the text-only baseline rather than consistently exceed it.

The pattern on LifeSnaps also highlights the value of chart-based visual encoding for continuous sensor trajectories. Stress and anxiety-related signals often involve gradual and interacting changes rather than discrete symbolic events. When those changes are represented visually, the VLM can exploit shape-based cues that are difficult to encode fully in text.

4.7.3 Results on MFAFY

MFAFY is the boundary case of the chapter, and its results are therefore especially informative. Unlike GLOBEM and LifeSnaps, MFAFY is centered on qualitative, non-cognitive educational trajectories rather than continuous sensor streams. As a result, the visual channel has less naturally continuous structure to exploit.

The pattern reflects this. Time-series baselines and Time-VLM remain weak, with OOD accuracy between 44% and 50%. LLM-Meta reaches 64.86% OOD accuracy and 64.14% F1. LE-Viz still improves OOD accuracy, reaching 65.43% with heatmaps and 66.57% with charts. However, the F1 advantage is less uniform: the chart variant reaches 62.62% OOD F1, which is below the 64.14% OOD F1 of LLM-Meta.

This does not undermine the multimodal argument. Instead, it clarifies its scope. Visual encoding appears to provide the largest gains when the underlying data contains continuous temporal dynamics that can be expressed as shape, slope, or spatial progression. When the features are more categorical or purely qualitative, the visual channel still helps on accuracy, but its advantage narrows. MFAFY therefore serves an important role in the chapter: it shows that multimodality is beneficial, but not equally so for every feature type.

4.7.4 ID-to-OOD stability

Another useful way to read Table 4.2 is through the size of the ID-to-OOD drop. On GLOBEM, the chart-based LE-Viz model drops from 79.37% ID accuracy to 72.86% OOD accuracy, a gap of 6.51 points. This is smaller than what would be expected if the model were simply overfitting to source-specific patterns and then collapsing under shift. Similar patterns appear on LifeSnaps and MFAFY. The multimodal models do not remove the shift problem, but they reduce its severity relative to weaker baselines.

That observation matters because the dissertation is not pursuing accuracy alone. It is pursuing representations that hold up when the data distribution changes. LE-Viz improves performance in that exact regime.

4.8 Ablation evidence: why LE-Viz works

The main results establish that LE-Viz works. The ablations clarify *why*. This is one of the most important parts of the chapter because it converts the multimodal gain from an empirical fact into a more interpretable representational claim.

4.8.1 Not all VLMs benefit equally

Table 4.3 compares three VLMs under the same optimal input configuration on GLOBEM. LLaVA-NeXT 7B performs best, reaching 72.86% OOD accuracy and 74.92% OOD F1. MiniGPT-4 13B reaches 66.68% OOD accuracy and 64.00% F1, while mPLUG-Owl2 8.2B reaches 65.74% OOD accuracy and 63.77% F1. Both of the latter models underperform the text-only LLM-Meta baseline on OOD F1.

This is a critical result because it rules out an overly simple reading of the chapter. The lesson is not that visual encoding automatically helps every VLM.

Table 4.3: Effect of VLM architecture on GLOBEM using the same LE-Viz input configuration (Meta-Narrative + interleaved charts). The result shows that architecture matters: better visual priors and cross-modal alignment are more important than nominal parameter count alone [21].

	ID		OOD	
Model	Acc (%)	F1 (%)	Acc (%)	F1 (%)
LLaVA-NeXT 7B	**79.37**	**82.03**	**72.86**	**74.92**
MiniGPT-4 13B	73.09	70.87	66.68	64.00
mPLUG-Owl2 8.2B	70.85	65.97	65.74	63.77
Text-only LLM-Meta	73.99	73.87	67.40	69.40

Rather, the benefit depends on whether the visual encoder and multimodal connector are well matched to structured, chart-like temporal inputs. LE-Viz therefore makes a more nuanced claim: *multimodality helps when the representation and the architecture are aligned*. The result also suggests that the quality of pre-trained visual primitives matters. Low-level visual features learned from natural images appear to transfer to temporal charts, but they do so unevenly across encoders.

4.8.2 Spatial organization matters, not just information quantity

Perhaps the strongest ablation in the chapter is the comparison between interleaved and combined visual organization. Both contain the same underlying visual information, but they arrange it differently. Table 4.4 shows that on GLOBEM, interleaved charts achieve 72.86% OOD accuracy and 74.92% F1, whereas a single combined chart image drops to 62.48% OOD accuracy and 64.18% F1. Heatmaps show the same directional pattern, though with a smaller gap.

This matters enormously for the interpretation of LE-Viz. If the multimodal gain came simply from "more information," then changing the spatial arrangement of identical information should not produce such a large difference. The fact that it does indicates that the gain is driven by *representational structure*. In the interleaved

Table 4.4: Effect of visual organization on GLOBEM. The interleaved format preserves local text-vision correspondences and yields substantially better generalization than placing the same visual information into a single combined image [21].

	ID				OOD			
Method	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
LE-Viz (Interleaved Chart)	**79.37**	**83.33**	**80.77**	**82.03**	**72.86**	**75.58**	**74.28**	**74.92**
LE-Viz (Interleaved Heatmap)	76.68	72.83	71.28	72.04	70.98	73.19	73.13	73.16
LE-Viz (Single Combined Chart)	67.26	76.27	66.67	71.15	62.48	62.16	66.34	64.18
LE-Viz (Single Combined Heatmap)	72.65	78.26	71.43	74.69	66.73	65.09	67.42	66.23

format, each feature's textual and visual evidence appears locally together, which supports feature-specific cross-modal grounding. In the combined format, those correspondences must be learned across longer distances and are therefore less reliable under shift.

At the dissertation level, this is the moment where the chapter's central argument becomes most precise: *how the information is spatially organized matters, not just whether the information is present*. LE-Viz succeeds because it preserves local structure in a form the VLM can exploit.

4.8.3 The visual channel is weaker alone, stronger in combination

Table 4.5 presents a controlled component analysis on LifeSnaps. Across all three VLMs, text-only outperforms vision-only, but multimodal text+vision outperforms both. For LLaVA-NeXT 7B on OOD evaluation, text-only reaches 67.0% accuracy, vision-only reaches 57.3%, and the multimodal combination reaches 71.9%.

This result reveals an important asymmetry. The textual Meta-Narrative remains the stronger single modality because it already carries interpreted semantic content. The visual channel by itself is weaker, which means LE-Viz is not a story about vision replacing language. It is a story about vision providing comple-

Table 4.5: Controlled component analysis on LifeSnaps. Text-only is the stronger single modality, but text+vision consistently outperforms both text-only and vision-only across all three VLMs, indicating genuine complementarity rather than simple modality substitution [21].

VLM	ID (Acc / F1)			OOD (Acc / F1)		
	Text-only	Vision-only	Text+Vision	Text-only	Vision-only	Text+Vision
LLaVA-NeXT 7B	65.5 / 68.9	58.4 / 61.3	**82.4 / 80.0**	67.0 / 72.8	57.3 / 59.9	**71.9 / 76.4**
MiniGPT-4 13B	63.1 / 62.4	60.8 / 63.6	**73.2 / 72.2**	61.7 / 61.0	56.8 / 61.0	**68.4 / 71.2**
mPLUG-Owl2 8.2B	64.5 / 63.8	58.3 / 60.8	**70.3 / 69.8**	64.3 / 63.6	55.8 / 58.4	**69.8 / 71.4**

mentary structural information that language alone does not fully capture. That complementarity is exactly what a multimodal LE framework should look like.

4.8.4 A compact synthesis of the ablations

The ablations can be summarized by three takeaways:

1. **Architecture matters.** Purpose-designed visual representations help only when the VLM can process them effectively.

2. **Organization matters.** Interleaving text and visual evidence yields much stronger OOD transfer than aggregating the same visual information into a distant combined block.

3. **Complementarity matters.** Text remains the stronger semantic modality, but visual encoding adds structure that text-only modeling leaves under-preserved.

These points collectively support the claim that LE-Viz's gains come from structure-preserving multimodal design rather than from indiscriminate use of more modalities.

4.9 Why chart-based visual encoding helps more than heatmaps

A consistent pattern in Table 4.2 is that chart-based LE-Viz outperforms heatmap-based LE-Viz across all three datasets. This pattern deserves interpretation because it speaks directly to how VLMs read temporal graphics.

Line charts preserve several properties with unusual clarity: absolute direction of change, slope, curvature, sudden breakpoints, and relative stability. For continuous sensor-driven trajectories, those cues are exactly the kinds of patterns the model needs. Heatmap chains preserve temporal order as well, but they convert much of the value geometry into color intensity rather than explicit position. This can be advantageous for some kinds of variation, but it may provide weaker cues for fine-grained directional reasoning.

The results suggest that in the LE settings considered here, the VLM benefits more from trajectory geometry than from intensity-coded progression alone. This is especially plausible for GLOBEM and LifeSnaps, where shape-based changes in continuous features such as sleep, activity, or stress are likely to be informative. It also helps explain why the heatmap variant remains strong, but not strongest: it preserves structure, though with a different balance of visual emphasis.

At the same time, the chapter does not claim that charts are universally optimal. LE-Viz tests only two visual formats, both hand-designed. The broader implication is that visualization design itself is a modeling choice. Once LE data is understood as something that can be rendered into different spatial grammars, the question of *which* grammar best supports generalization becomes part of the learning problem.

4.10 Why generic multimodal time-series VLMs are not enough

One of the most revealing empirical comparisons in the chapter is LE-Viz versus Time-VLM. Time-VLM is already multimodal in the broad sense, yet it performs at or below the dedicated time-series baselines on all three datasets. This is important because it shows that the multimodal gain in LE-Viz is not explained by the mere presence of images or vision-language components.

The underlying reason is representational mismatch. Time-VLM was designed for more conventional numerical forecasting settings. LE data is different: it is heterogeneous, behaviorally contextual, and often includes both sensor-derived and subjective signals. LE-Viz explicitly addresses that heterogeneity by combining Meta-Narratives with feature-specific visual encodings and by using an interleaved format that grounds each visual cue in local textual context. Time-VLM does not solve that representational problem, so its multimodal machinery does not translate into comparable gains.

This comparison strengthens the dissertation's larger methodological point. In LE forecasting, the main challenge is not simply to choose a more powerful model family. It is to design a representation that respects the structure of the data. LE-Viz succeeds because it is representationally specific.

4.11 What LE-Viz contributes to the dissertation arc

LE-Viz occupies a crucial place in the dissertation because it changes the representation story in two ways.

First, it demonstrates that the strongest text-only stage of the dissertation still left meaningful structure underused. This is not a critique of ConText-LE. Rather, it is the logical continuation of that chapter's success. Once Meta-Narratives proved

that semantic contextualization matters, the next question became whether the structural information lost in serialization could also be recovered. LE-Viz answers that question positively.

Second, LE-Viz shows that multimodality is not merely an optional enhancement. For LE data, it is a representational correction. The visual channel restores aspects of temporal locality, grouping, and trajectory geometry that the text channel alone cannot preserve with equal fidelity. The results across GLOBEM and LifeSnaps make that claim strongly, while MFAFY clarifies the limits and conditions under which the visual advantage is largest.

The chapter therefore moves the dissertation from *semantic generalization* to *structure-preserving multimodal generalization*. That transition is central to the dissertation's overall intellectual arc.

4.12 Limitations and practical constraints

Although LE-Viz substantially improves cross-distribution forecasting, it is not the final answer of the dissertation. Several limitations remain visible even within its strongest results.

4.12.1 Computational cost

The first limitation is cost. LE-Viz is heavier than the text-only systems in the previous chapter. It requires visual rendering in addition to textual narrative generation, larger multimodal inputs, and VLM fine-tuning on high-memory hardware. The reported implementation uses 8 A40 GPUs, which already signals a different computational regime from lightweight text-only modeling [21]. For a dissertation concerned with reliable and practical LE modeling, this matters. Better

representation is valuable, but not if it depends on fragile or excessively expensive adaptation.

4.12.2 Dependence on external generation components

LE-Viz also inherits some of the pipeline complexity of the prior narrative-based work. GPT-4o is used for Meta-Narrative generation, for target narrative creation during training, and for extraction of final predictions from generated outputs. These choices are methodologically effective, but they add cost and reproducibility concerns. The chapter therefore treats them honestly as part of the current solution rather than hiding them behind the model itself.

4.12.3 Data-type sensitivity

The MFAFY results show that the visual channel is not equally beneficial for every kind of LE data. When the underlying features are more categorical, qualitative, or less naturally continuous, visual encoding still helps on some metrics but yields smaller gains overall. This is an important boundary condition. The chapter does not argue that every LE task should automatically be rendered visually in the same way. Instead, it argues that visual encoding is most effective when the data contains structure that benefits from spatial representation.

4.12.4 Modality imbalance and overfitting

The component analysis shows that text remains the stronger standalone modality. That is useful, but it also signals a risk: a multimodal model may drift toward text dominance or vision dominance depending on which modality is easier to exploit during source-distribution training. More broadly, end-to-end multimodal adaptation in small-data settings always carries the danger of learning

modality-specific shortcuts rather than genuinely distribution-invariant structure. LE-Viz improves generalization substantially, but it does not fully eliminate this problem.

These limitations are not side notes. They are exactly what prevent this chapter from being the endpoint of the dissertation. LE-Viz establishes the necessity of multimodality. It also makes clear that multimodality needs stronger discipline if the goal is reliability under shift.

4.13 Chapter synthesis

The central contribution of this chapter can be summarized in one sentence: *LE-Viz shows that the next major gain in LE forecasting comes not from abandoning narrative language modeling, but from complementing it with structure-preserving visual encoding.*

More concretely, this chapter contributes the following to the dissertation:

1. it identifies the representational bottleneck of text-only LE modeling as the loss of two-dimensional Feature $\times$ Time structure;
2. it motivates visual encoding through principles of visuospatial reasoning rather than through architecture novelty alone;
3. it presents LE-Viz as a multimodal framework that pairs Meta-Narratives with charts or heatmap chains and performs prospective narrative generation with a VLM;
4. it demonstrates substantial OOD gains over time-series baselines, text-only LLM baselines, and a generic VLM baseline;

5. and it shows through ablations that architecture choice, local cross-modal organization, and text-vision complementarity are the real mechanisms behind the observed gains.

At the dissertation-question level, this chapter directly addresses **RQ4** and provides strong empirical support for **H4**. The results show that text-only serialization loses predictive structural information and that multimodal visual–text representation can recover that structure in ways that improve cross-distribution transfer.

The chapter also provides transitional evidence for **RQ5**. While LE-Viz improves generalization, it still exhibits computational burden, modality imbalance, and potential source-specific co-adaptation. These limitations motivate the constrained-learning claim in **H5**, which is taken up explicitly in the next chapter.

In the dissertation arc, this chapter therefore performs a dual role. It validates multimodality as a necessary correction to representational collapse, and it reveals why successful multimodal systems still require stronger constraints for reliability under shift.

4.14 Bridge to the Next Chapter

This chapter has shown that visual encoding can recover structure that text-only representations lose. By pairing Meta-Narratives with charts or heatmap chains, LE-Viz provides a more complete view of the same underlying trajectory and achieves markedly stronger cross-distribution forecasting than either time-series-only or text-only approaches. The evidence is strongest on datasets where the underlying signals contain continuous temporal dynamics, and the ablations show that the gains come from purposeful multimodal design rather than from

indiscriminate architectural complexity [21].

At the same time, the chapter has also exposed the remaining problem. The multimodal gain in LE-Viz is achieved through a relatively heavy adaptation pipeline that still leaves room for modality imbalance, over-specialization, and high computational cost. In other words, multimodality has become necessary, but it has not yet become sufficiently constrained. The next chapter addresses that issue by asking how complementary multimodal views can be preserved while reducing the freedom of the model to co-adapt in source-specific ways.

Chapter 5

PRISM: Frozen Multimodal Constraints for Generalizable and Reliable Longitudinal Experiential Modeling

5.1 Introduction

The previous chapter established that VLM-based modeling substantially improves LE forecasting over both traditional time-series models and text-only large language model (LLM) baselines. By pairing contextual *Meta-Narratives* with structure-preserving visual encodings, LE-Viz showed that multimodal representations recover temporal locality, cross-feature grouping, and trajectory geometry that are weakened when LE data is serialized into text alone [21, 75]. Those gains were large and consistent. The LE-Viz stage also surfaced a deeper point that becomes decisive in this chapter: *multimodality by itself is not the final answer*. A model may ingest multiple modalities and still overfit to source-distribution shortcuts, suppress one modality during adaptation, or incur an impractical computational burden when data are limited and distribution shift is severe [21, 76–78].

This chapter develops the final framework of the dissertation from that tension. Text-only modeling is structurally incomplete, and unrestricted end-to-end multimodal adaptation remains unstable and expensive. Under these conditions, the central question changes. The problem is no longer whether multiple views

of the same trajectory are useful. The previous chapter already established that they are. The problem becomes *how those views should be constrained so that they promote transfer instead of source-specific fit*. PRISM is proposed as the answer to that question.

PRISM, short for **P**rospective **R**easoning through **I**ntegrated **S**pectral-temporal **M**ultimodal learning, is the culminating technical contribution of this dissertation. It is motivated by a simple but consequential observation: cross-distribution failure in LE modeling is often not a consequence of insufficient model size, but of insufficient representational discipline during training. A prediction loss alone permits the model to discover source-specific sufficient statistics that perform well on the training distribution without encoding the deeper behavioral structure required for generalization. Likewise, a fine-tuned generative model may produce plausible narratives while gradually adapting its own notion of coherence to the source distribution, thereby weakening its transfer value [75, 79, 80].

PRISM addresses this problem through two intertwined commitments. First, it decomposes each behavioral trajectory into three functionally complementary views: a *temporal measurement stream*, a *spectral dynamics stream*, and a *semantic interpretation stream*. Second, it constrains the shared multimodal representation with a *frozen language prior*. In this formulation, language is not only the final reporting interface. It becomes a fixed semantic test that the learned representation must satisfy. The prediction objective enforces discriminative structure, while the frozen-language narrative objective enforces semantic coherence that cannot drift with the training distribution [21, 75].

This chapter serves several purposes simultaneously. It completes the dissertation arc technically, because it integrates the strongest insights from the earlier stages of the dissertation. It also completes the arc conceptually, because it re-

frames generalization as a problem of *representation under constraint*, not only model choice. The chapter therefore moves carefully from motivation to mechanism to evidence. It begins by explaining why the LE-Viz stage, despite its strong performance, still leaves a generalization gap. It then formalizes the PRISM problem setting, details the three-stream architecture and directed fusion mechanism, introduces the dual-path training formulation with homoscedastic task weighting, and presents detailed experimental evidence across GLOBEM, LifeSnaps, and MFAFY. The chapter concludes by interpreting why PRISM works, what it reveals about distribution-invariant reasoning in LE data, and why it stands as the strongest answer offered by this dissertation to the problem of reliable cross-distribution behavioral forecasting.

5.2 From LE-Viz to PRISM: Why Multimodality Was Necessary but Not Sufficient

The transition from Chapter 4 to the present chapter is a continuation of the same representational argument. PRISM builds directly on what LE-Viz made visible. That chapter showed that text-only representations remain bounded by serialization: even when a strong Meta-Narrative summarizes the trajectory semantically, the original Feature $\times$ Time organization is still compressed into a one-dimensional token sequence. Visual encoding helped because it restored aspects of structure that language alone could not preserve well [21]. However, once multimodality is introduced, a second representational challenge emerges.

A fine-tuned VLM is still free to solve the source task by whichever internal shortcut is easiest. In small or moderate behavioral datasets, that often means leaning excessively on whichever modality is easiest to optimize under the source

distribution. A model may appear multimodal at the input level while behaving unimodally or weakly multimodally in practice [76,77]. In addition, full or heavy multimodal adaptation is expensive, and the gains may come with a sizeable ID to OOD gap, suggesting that some of the acquired representation remains tied to the source environment instead of robust behavioral abstractions [21].

These observations motivate a more disciplined principle for the final stage of the dissertation:

> *Generalization does not improve from adding more modalities alone. It improves when complementary modalities are organized so that the learned representation must satisfy constraints that are harder to fulfill through source-specific shortcuts.*

PRISM is built around this principle. It retains the insight that multiple views of the same trajectory are valuable, but it changes the training logic. Instead of allowing the full multimodal system to co-adapt freely, PRISM keeps the large pretrained language and CLIP backbones frozen and learns only lightweight task-specific components and modality bridges. This matters not only for efficiency. It matters because freezing prevents the semantic criterion itself from drifting toward the training distribution. The system must therefore discover a representation that is useful for prediction *and* sufficiently structured to support coherent prospective narrative generation from a language model whose internal prior remains fixed [75,81].

5.3 Representational Diagnosis: Why Cross-Distribution Failure Persists

Let a behavioral trajectory for participant i be denoted by

$$\mathbf{X}_i \in \mathbb{R}^{T \times d}, \tag{5.1}$$

where T is the number of time steps and d is the number of behavioral features. The task is to predict an outcome $y_i \in \mathcal{Y}$ for a held-out future state or evaluation window. In the cross-distribution setting, training examples come from a source distribution $\mathcal{D}_T$, while the model is evaluated on a target distribution $\mathcal{D}_{T'}$ with

$$P_{\mathcal{D}_T}(\mathbf{X}, y) \neq P_{\mathcal{D}_{T'}}(\mathbf{X}, y). \tag{5.2}$$

The challenge is not only that values shift. Behavioral meaning itself is context-dependent. A decline in mobility, an increase in sleep duration, or a change in communication frequency may imply different future outcomes depending on the broader temporal pattern and participant context [82,83]. Thus, a model that learns a brittle decision boundary over raw measurements may fit the source distribution while failing to encode the deeper structures that remain meaningful under shift.

The dissertation's earlier chapters progressively identified the main sources of this failure. Chapter 2 showed that classical ML and DL systems flatten or aggregate trajectories in ways that weaken temporal context. Chapter 3 showed that LLMs help because narrative representations provide semantic framing, contextual reasoning, and denser generative supervision. Chapter 4 showed that even strong textualization remains structurally incomplete because the native organization of LE data is two-dimensional. PRISM synthesizes these lessons into a final diagnosis:

robust LE forecasting requires a representation that preserves at least three distinct information types simultaneously:

1. **Measurement information**: the literal observed values and their cross-feature correlations.

2. **Dynamics information**: how the trajectory changes across temporal scales, including non-stationary disruptions.

3. **Semantic information**: what those patterns mean in behavioral context.

No single representation is sufficient for all three. Raw sequences preserve measurements but not semantic interpretation. Textual meta-narratives preserve context and interpretation but abstract away from exact numeric detail and some local temporal geometry. Spectral representations capture change structure but are lossy with respect to exact magnitudes and contextual meaning. PRISM therefore treats complementarity as a design requirement, not a convenience.

5.4 Problem Formulation

Formally, let

$$\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N} \tag{5.3}$$

denote a dataset of longitudinal behavioral trajectories. Each example additionally has an associated meta-narrative $\mathbf{M}_i$, generated in preprocessing following the narrative protocol introduced in ConTeXt-LE [75]. The cross-distribution generalization problem is to learn a forecasting function

$$f : \mathbb{R}^{T \times d} \to \mathcal{Y} \tag{5.4}$$

that minimizes error on $\mathcal{D}_{T'}$ without access to the target distribution during training.

PRISM reframes this task as a constrained multimodal representation problem. Instead of learning directly from a single input view, each trajectory is decomposed into three streams:

$$\mathbf{X}_i \to \left(\mathbf{x}_i^{\text{temp}}, \mathbf{x}_i^{\text{spec}}, \mathbf{x}_i^{\text{sem}}\right), \tag{5.5}$$

where $\mathbf{x}_i^{\text{temp}}$ is the temporal measurement view, $\mathbf{x}_i^{\text{spec}}$ is the spectral dynamics view, and $\mathbf{x}_i^{\text{sem}}$ is the semantic narrative view. These streams are encoded separately, fused through directed cross-modal interaction, and then constrained by two output objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda\, \mathcal{L}_{\text{coh}}, \tag{5.6}$$

where $\mathcal{L}_{\text{pred}}$ enforces discriminative forecasting structure and $\mathcal{L}_{\text{coh}}$ enforces coherence with a frozen language prior. In the implemented PRISM model, this balance is learned dynamically through homoscedastic uncertainty weighting instead of a fixed manual λ [84].

The conceptual claim behind this formulation is central to the chapter: source-specific shortcut learning often satisfies the prediction loss because a coarse label can be correct for the wrong reason. By contrast, producing a coherent prospective narrative from a frozen language model is a stricter test. The representation must carry enough structured information to support semantically plausible forecasting, not only class separation.

5.5 Overview of the PRISM Framework

PRISM decomposes each behavioral trajectory into three functionally complementary information streams, processes each through a dedicated encoder, fuses the resulting token-level representations through directed pairwise cross-attention and gated aggregation, and trains the shared representation under a dual-objective formulation. Figure 5.1 shows the overall system.

A concise summary of the architecture is given in Table 5.1. This table is useful because PRISM is not just a multimodal model in the ordinary sense. All three streams arise from the same underlying trajectory, but each preserves a different information dimension through a deliberately chosen lossy transformation.

5.6 Functionally Complementary Stream Decomposition

5.6.1 Temporal measurement stream

The temporal stream preserves the raw multivariate trajectory in its most direct behavioral form. Given $\mathbf{X}_i \in \mathbb{R}^{T \times d}$, the input is projected into a latent space by

$$\mathbf{H}^{(0)} = \mathbf{X}_i \mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}, \tag{5.7}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d \times d_{\text{model}}}$ and $d_{\text{model}} = 256$. Sinusoidal positional encodings are added to retain order:

$$\mathbf{Z}^{(0)} = \mathbf{H}^{(0)} + \mathbf{P}. \tag{5.8}$$

The sequence then passes through $L = 2$ transformer encoder layers:

$$\mathbf{Z}^{(l)} = \text{TransformerLayer}(\mathbf{Z}^{(l-1)}), \qquad l = 1, \ldots, L, \tag{5.9}$$

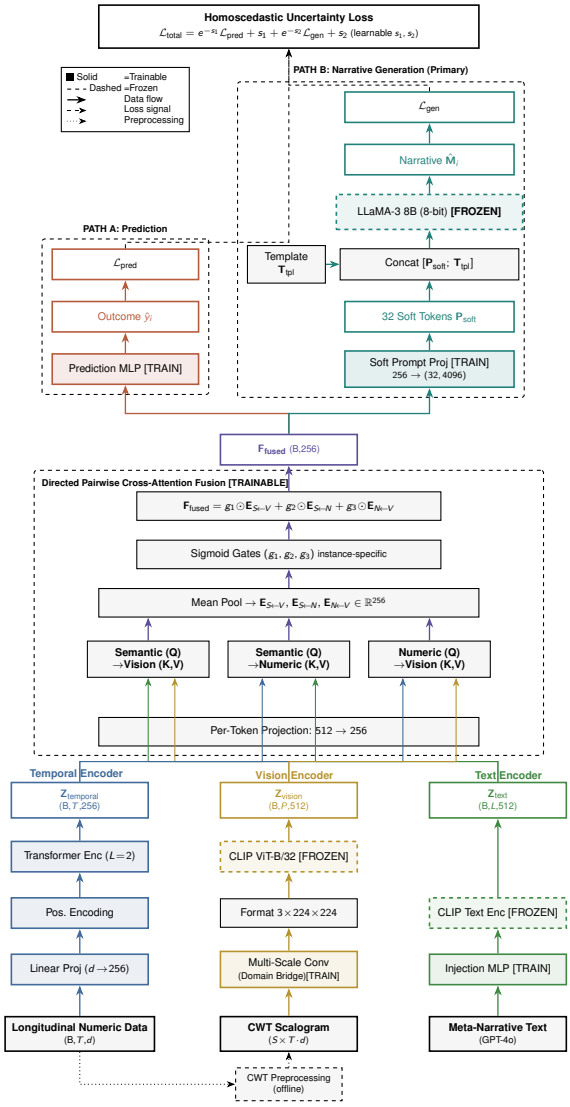


Figure 5.1: The PRISM framework. Longitudinal behavioral trajectories are decomposed into temporal measurement, spectral dynamics, and semantic interpretation streams. The temporal stream is processed by a trainable transformer encoder, while the visual and textual streams pass through frozen CLIP encoders with trainable input-side adaptation. Directed pairwise cross-attention produces enriched cross-modal representations, which are aggregated through learned instance-specific gates into a fused embedding. The fused representation supports both direct prediction and prospective narrative generation through soft prompting of a frozen LLaMA-3 decoder. Adapted from the PRISM paper source used to build this chapter [22].

Table 5.1: Component-level summary of PRISM. The key architectural idea is that each stream foregrounds one information type while relying on the others to recover what it discards.

Stream	**Input form**	**Encoder**	**Frozen?**	**Primary information preserved**
Temporal measurement	Raw multivariate sequence $\mathbf{X}_i \in \mathbb{R}^{T \times d}$	Lightweight transformer encoder	No	Exact feature values, cross-feature correlations, local volatility, ordered measurement evidence
Spectral dynamics	CWT scalogram derived from concatenated trajectory	CLIP vision encoder with trainable multi-scale input bridge	Yes (backbone)	Time-frequency localization, disruption timing, multi-scale periodicity, non-stationary dynamics
Semantic interpretation	Meta-Narrative text $\mathbf{M}_i$	CLIP text encoder with trainable injection MLP	Yes (backbone)	Contextual behavioral meaning, high-level interpretation, semantic abstraction of trajectory patterns
Fusion and output	Token sequences from all streams	Directed pairwise cross-attention, gated aggregation, prediction head, soft-prompt projection to frozen LLaMA-3	Mixed	Cross-modal retrieval, instance-specific pathway weighting, discriminative forecasting, narrative coherence

producing

$$\mathbf{Z}_{\text{temporal}} = \mathbf{Z}^{(L)} \in \mathbb{R}^{T \times d_{\text{model}}}. \tag{5.10}$$

This stream retains the measurement-level evidence that later semantic and spectral views abstract away from. It is the only stream in PRISM whose backbone

is fully trainable from scratch. That choice is intentional. The raw temporal stream must remain flexible enough to model dataset-specific feature interactions and local sequence behavior. At the same time, its comparatively smaller contribution in later ablations will prove informative: exact measurements remain necessary, but they are not sufficient to drive robust transfer on their own.

5.6.2 Spectral dynamics stream

The spectral stream is built around the observation that many important behavioral changes are *non-stationary*. For visualizing temporal behavior, one can use line charts or frequency-domain summaries such as the Fast Fourier Transform (FFT). Line charts are useful for local shape inspection, and FFT captures global periodic content. However, FFT does not indicate when a frequency-specific event occurs. In LE forecasting, that timing is often important: a disruption localized to one week, a sudden loss of regularity in sleep, or a brief but meaningful volatility burst in phone activity can have different implications from the same global frequency content spread evenly across time. PRISM therefore uses the Continuous Wavelet Transform (CWT), which preserves joint time-frequency localization [85, 86].

For a univariate signal $x(t)$, the CWT with Morlet wavelet ψ is

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt, \tag{5.11}$$

where $a > 0$ is the scale parameter and b is the temporal translation. This formulation is preferred here because it captures both *which* scale is active and *when* it is active. In PRISM, the d features are concatenated along the temporal axis to form a univariate signal of length $T \cdot d$, from which a scalogram is computed

using logarithmically spaced scales. The resulting representation captures the timing and scale of behavioral changes.

This design is conceptually important. The spectral stream does not attempt to preserve exact feature identity. That responsibility belongs to the temporal stream. Instead, the spectral stream foregrounds *how the overall behavioral structure changes across time and scale*. To bridge the domain gap between behavioral scalograms and CLIP's natural-image pretraining, PRISM introduces trainable multi-scale convolutions that transform the scalogram into a three-channel image suited to the frozen CLIP vision encoder:

$$\mathbf{Z}_{\text{vision}} = \text{CLIP}_{\text{vision}}\big(\text{Conv}_\theta(\mathbf{S}_i)\big) \in \mathbb{R}^{P \times d_{\text{clip}}}, \tag{5.12}$$

where $\mathbf{S}_i$ is the normalized scalogram, P is the number of patch tokens, and $d_{\text{clip}} = 512$.

The domain bridge is a noteworthy architectural decision. Rather than fine-tuning the entire vision backbone on small behavioral datasets, PRISM keeps the CLIP encoder frozen and learns only how to present the scalogram to it. This reduces overfitting risk while still exploiting a powerful pretrained visual feature extractor [87].

5.6.3 Semantic interpretation stream

The semantic stream carries the strongest connection to Chapter 3. Each trajectory is associated with a contextual Meta-Narrative generated using the ConTeXt-LE pipeline. This narrative is not a literal transcription of raw values. It is an interpretation of the trajectory that situates observable patterns within a broader behavioral frame [75]. In PRISM, this narrative is encoded through a

frozen CLIP text encoder preceded by a trainable injection MLP:

$$\mathbf{Z}_{\text{text}} = \text{CLIP}_{\text{text}}\big(\text{MLP}_{\phi}(\text{Embed}(\mathbf{M}_i))\big) \in \mathbb{R}^{L \times d_{\text{clip}}}, \tag{5.13}$$

where L is the token length of the narrative.

The motivation for this stream is subtle but crucial. Language does not only summarize behavior. It encodes what the model should regard as behaviorally meaningful. A simultaneous drop in activity and shift in sleep, for example, is not important solely because the values changed. It becomes predictive because of the context in which such a configuration typically matters. The semantic stream therefore preserves an interpretive prior that the raw temporal stream and spectral stream cannot generate by themselves.

Using CLIP instead of a standard language-only encoder for this branch also serves the fusion strategy. Because the CLIP text and CLIP vision encoders were pretrained jointly, their latent spaces already carry some degree of alignment. This makes the subsequent text-vision attention in PRISM more grounded than an arbitrary pairing of unrelated encoders.

5.6.4 Why these streams are complementary

PRISM's decomposition is best understood as a hierarchy from concrete evidence to abstract meaning:

- the temporal stream preserves *what was measured and when*,
- the spectral stream preserves *how the structure of behavior changed across scales*,
- the semantic stream preserves *what those changes mean in context*.

Each transformation is lossy in a different way. Spectral analysis discards exact magnitudes and some feature identity; narrative abstraction discards low-level numeric precision; raw temporal encoding does not carry enough high-level interpretive structure. The point of PRISM is therefore not redundancy. It is to ensure that the shared representation must reconcile information that is difficult to satisfy through a single shortcut.

5.7 Directed Cross-Modal Fusion

A standard multimodal baseline might concatenate the pooled outputs of the three streams and pass them through an MLP. PRISM rejects that design for a principled reason. The three streams are not functionally symmetric. Semantic tokens should attend to spectral patches and temporal positions in order to retrieve evidence relevant to interpretation. Temporal tokens should attend to spectral patches in order to contextualize local measurements in a broader dynamical pattern. These are directed relations, not just undifferentiated feature combinations.

The CLIP-derived sequences are first projected to the temporal model dimension. PRISM then applies three directed multi-head cross-attention blocks:

$$\mathbf{H}_{S\leftarrow V} = \mathrm{MHA}(\mathbf{Z}^{\downarrow}_{\text{text}}, \mathbf{Z}^{\downarrow}_{\text{vision}}, \mathbf{Z}^{\downarrow}_{\text{vision}}), \tag{5.14}$$

$$\mathbf{H}_{S\leftarrow N} = \mathrm{MHA}(\mathbf{Z}^{\downarrow}_{\text{text}}, \mathbf{Z}_{\text{temporal}}, \mathbf{Z}_{\text{temporal}}), \tag{5.15}$$

$$\mathbf{H}_{N\leftarrow V} = \mathrm{MHA}(\mathbf{Z}_{\text{temporal}}, \mathbf{Z}^{\downarrow}_{\text{vision}}, \mathbf{Z}^{\downarrow}_{\text{vision}}), \tag{5.16}$$

where S, N, and V denote semantic, numeric, and vision streams respectively.

Each enriched sequence is then mean-pooled:

$$\mathbf{E}_{S\leftarrow V} = \text{MeanPool}(\mathbf{H}_{S\leftarrow V}), \quad \mathbf{E}_{S\leftarrow N} = \text{MeanPool}(\mathbf{H}_{S\leftarrow N}), \quad \mathbf{E}_{N\leftarrow V} = \text{MeanPool}(\mathbf{H}_{N\leftarrow V}). \tag{5.17}$$

The directionality of these interactions embodies a behavioral hypothesis. Information should flow from evidence into interpretation. Semantic tokens query temporal and spectral evidence to ground their interpretive content. Temporal tokens query spectral evidence to situate raw change patterns within a multi-scale dynamics view. PRISM does not rely on a symmetric attention soup in which all modalities mingle indiscriminately.

After directed interaction, the pooled vectors are aggregated through learned instance-specific sigmoid gates:

$$\mathbf{g} = \sigma\big(\mathbf{W}_g[\mathbf{E}_{S\leftarrow V}; \mathbf{E}_{S\leftarrow N}; \mathbf{E}_{N\leftarrow V}] + \mathbf{b}_g\big)\,, \tag{5.18}$$

$$\mathbf{F}_{\text{fused}} = g_1 \odot \mathbf{E}_{S\leftarrow V} + g_2 \odot \mathbf{E}_{S\leftarrow N} + g_3 \odot \mathbf{E}_{N\leftarrow V}. \tag{5.19}$$

These gates provide two advantages. First, they allow the model to suppress unreliable pathways for a specific participant. Second, they make the architecture interpretable at the pathway level, because they reveal which cross-modal relations the model relies on most strongly. Although the PRISM paper's gate-weight analysis was not fully finalized for every dataset, the modality ablations later in this chapter reveal an interpretable ordering that strongly corroborates the gating rationale.

5.8 Frozen Backbone Design and Why It Matters

A defining characteristic of PRISM is that its major foundation backbones remain frozen. Let Θ_{FM} denote the parameters of the CLIP and LLaMA components, and let Θ_{task} denote the trainable parameters in the temporal encoder, input-side adaptation modules, fusion layers, gates, prediction head, and prompt projection. PRISM enforces

$$\Theta_{\text{FM}}^{\star} = \Theta_{\text{FM}}^{(0)}, \tag{5.20}$$

while optimizing only

$$\Theta_{\text{task}}^{\star} = \arg\min_{\Theta_{\text{task}}} \mathcal{L}(\Theta_{\text{task}}; \mathcal{D}_T). \tag{5.21}$$

This choice plays several roles simultaneously.

Efficiency. Compared with LE-Viz, which fine-tuned a 7B VLM with LoRA and required substantially heavier multi-GPU training, PRISM is far lighter. The PRISM implementation runs on a single NVIDIA A40 GPU with the large backbones frozen, while LE-Viz relied on 8 NVIDIA A40 GPUs for multimodal fine-tuning [21,22]. In this sense, PRISM is not only more generalizable but also more practical.

Stability. Frozen backbones reduce the risk that foundation-model priors drift toward spurious source-distribution patterns. The semantic criterion remains anchored in broad pretraining instead of being reshaped by the relatively small behavioral training set.

Constraint strength. Most importantly, freezing changes the logic of learning. In standard multitask or multimodal fine-tuning, every component can adapt

together and potentially collude around source-specific solutions. In PRISM, the LLM's notion of narrative coherence cannot be updated during training. The representation must therefore move toward the fixed prior, not the other way around.

This is why freezing should be regarded as more than an efficiency trick. In PRISM, it is part of the generalization mechanism.

5.9 Dual-Path Learning: Prediction and Frozen-Language Coherence

5.9.1 Path A: direct outcome prediction

The fused representation supports a lightweight discriminative head:

$$\hat{y}_i = \mathrm{MLP}_{\mathrm{pred}}(\mathbf{F}_{\mathrm{fused}}), \tag{5.22}$$

with binary cross-entropy loss

$$\mathcal{L}_{\mathrm{pred}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log \sigma(\hat{y}_i) + (1-y_i)\log(1-\sigma(\hat{y}_i))\Big]. \tag{5.23}$$

This path provides the direct forecasting signal that ties the representation to the target label.

5.9.2 Path B: prospective narrative generation via soft prompting

The same fused representation is projected into k soft prompt tokens in the input embedding space of a frozen LLaMA-3 model:

$$\mathbf{P}_{\mathrm{soft}} = \mathrm{MLP}_{\mathrm{prompt}}(\mathbf{F}_{\mathrm{fused}}) \in \mathbb{R}^{k \times d_{\mathrm{llm}}}, \tag{5.24}$$

where $k = 32$ and $d_{\text{llm}} = 4096$. These tokens are concatenated with a fixed prompt template and fed to the frozen decoder:

$$\hat{\mathbf{M}}_i = \text{LLaMA}_{\text{frozen}}\big([\mathbf{P}_{\text{soft}}; \mathbf{T}_{\text{template}}]\big). \tag{5.25}$$

The generation loss is the standard next-token cross-entropy against a target prospective narrative $\mathbf{M}_i^{\text{target}}$:

$$\mathcal{L}_{\text{gen}} = -\frac{1}{|\mathbf{M}_i^{\text{target}}|} \sum_{j=1}^{|\mathbf{M}_i^{\text{target}}|} \log p(m_j \mid m_{<j}, \mathbf{P}_{\text{soft}}, \mathbf{T}_{\text{template}}). \tag{5.26}$$

The crucial point is that gradients from this loss do not update the LLaMA model itself. They flow backward only into the prompt projection, fusion layers, and trainable stream adapters. Thus the frozen LLM acts as a *semantic coherence test*. The representation is rewarded only when it is rich enough to support a plausible behavioral forecast under a language prior that remains fixed.

5.9.3 Why this is not ordinary multitask learning

PRISM superficially resembles multitask learning because it uses two losses on a shared representation. But the resemblance is incomplete. In ordinary multitask training, all task heads are trainable and may adapt jointly to the source distribution. In PRISM, one of the two tasks is defined by a frozen prior. The generation path therefore imposes a constraint whose criterion of correctness cannot drift during training.

This difference is fundamental. It makes PRISM closer in spirit to a constrained learning framework than to a standard auxiliary-task design. The prediction head enforces class discrimination. The frozen generation path enforces reasoning-

compatibility with a semantic prior external to the training distribution. The shared representation must satisfy both.

5.9.4 Homoscedastic task uncertainty weighting

Because the two losses operate on very different scales, PRISM uses learned homoscedastic task uncertainty parameters s_1 and s_2 [84]:

$$\mathcal{L}_{\text{total}} = e^{-s_1}\mathcal{L}_{\text{pred}} + s_1 + e^{-s_2}\mathcal{L}_{\text{gen}} + s_2. \tag{5.27}$$

The inverse-variance terms e^{-s} act as learned task weights. A larger learned s corresponds to greater uncertainty and thus lower effective influence. In practice, the converged values reveal that PRISM consistently assigns more weight to the generation path, especially on the sensor-rich datasets. This finding becomes empirically important later in the chapter.

5.9.5 Inference

PRISM includes two inference modes. **Path A** uses the prediction head directly. **Path B**, the primary inference mechanism, generates a prospective narrative from the frozen LLM and extracts the forecast label from that narrative using the ConTeXt-LE parsing procedure [75]. Reporting both paths is analytically valuable because it separates two questions:

1. How good is the learned representation when evaluated by a lightweight discriminative head?

2. How much additional value emerges when that same representation is forced through a reasoning-like generative bottleneck?

The answer, as the experiments show, is that the generative path is substantially stronger.

5.10 Experimental Setup

5.10.1 Datasets and evaluation protocol

PRISM is evaluated on three completed cross-distribution LE benchmarks spanning mental health, anxiety, and educational engagement. Table 5.2 summarizes the settings. All models are trained on the source distribution and evaluated on both an ID validation/test split and a held-out OOD test set. The primary metric is OOD accuracy, with precision, recall, and F1 reported as complementary measures.

Table 5.2: Datasets used to evaluate PRISM. The three benchmarks differ not only in domain but also in the type of distribution shift and the nature of the input signals, making them a strong test bed for the final framework.

Dataset	**Domain**	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{P}'\rvert$	$\lvert\mathcal{D}_T\rvert$	$\lvert\mathcal{D}_{T'}\rvert$	**Features**	**OOD shift**
GLOBEM	Mental health	344	317	2226	2023	15	Cross-institution/year
LifeSnaps	Anxiety	26	13	112	64	15	Cross-participant and cross-temporal
MFAFY	Education	61	35	610	350	15	Cross-year

GLOBEM is the central benchmark of the dissertation and remains one of the strongest stress tests for cross-distribution behavioral forecasting [12]. LifeSnaps provides a substantially smaller but sensor-rich anxiety forecasting setting [88]. MFAFY differs sharply from the other two because its inputs are primarily categorical self-reports instead of continuous sensor trajectories, making it an important boundary case for the spectral stream [15, 89].

5.10.2 Baselines

The evaluation compares PRISM against three baseline families:

1. **Time-series models**: PatchTST and iTransformer, representing strong numeric sequence baselines without language or vision components [48, 49].

2. **LLM models**: LLM-Complete, LLM-Summary, LLM-Clustered, and ConTeXt-LE, spanning increasingly sophisticated text representations from raw serialization to contextual Meta-Narrative modeling [15, 75, 89–91].

3. **VLM models**: Time-VLM, which uses FFT-based visual encoding and statistical text descriptions, and LE-Viz, implemented by fine-tuning LLaVA-NeXT 7B with LoRA on interleaved LE charts and narratives [21, 92, 93].

5.10.3 Implementation details

The temporal encoder uses a model dimension of 256 with two transformer layers and four attention heads. The frozen visual and textual backbones are CLIP ViT-B/32 encoders, while the frozen generative model is LLaMA-3 8B loaded in 8-bit form. The soft prompt projection outputs $k = 32$ tokens. Training uses AdamW with learning rate 10^{-4}, weight decay 10^{-2}, batch size 32, and early stopping on validation loss. CWT scalograms use a Morlet wavelet with logarithmically spaced scales. The full PRISM implementation runs on a single NVIDIA A40 GPU [22].

This setup highlights a practically meaningful point for the dissertation: the final framework is more reliable than the earlier VLM stage while also being more efficient to train.

5.11 Main Results

Table 5.3 presents the full comparative results across the three evaluated datasets. The table is large, but that scale is appropriate here because the final

chapter must demonstrate not only that PRISM works, but that it improves over all prior representational stages in a coherent way.

5.11.1 Dataset-wise interpretation

GLOBEM. PRISM reaches 79.93% OOD accuracy on GLOBEM. This is more than 27 percentage points above the original GLOBEM benchmark state of the art at 52.8% and 12.53 points above the strongest prior text-only result from ConTeXt-LE at 67.40% [12,75]. It also improves by 7.07 points over LE-Viz and by 28.50 points over Time-VLM. The result is especially significant because GLOBEM is both the most established benchmark in the dissertation and one of the most difficult distribution-shift settings. The gain therefore cannot be dismissed as a quirk of a narrow task. It indicates that PRISM's combination of constrained multimodal fusion and frozen-language coherence materially changes the quality of the learned representation.

LifeSnaps. On LifeSnaps, PRISM reaches 81.25% OOD accuracy. The strongest earlier text-only baseline, ConTeXt-LE, achieved 67.19%, while LE-Viz reached 71.88%. Thus PRISM improves over the strongest text-only baseline by 14.06 points and over the strong multimodal baseline by 9.37 points. This matters because LifeSnaps is smaller than GLOBEM and therefore more vulnerable to overfitting. PRISM's success here supports the claim that the frozen design is not only theoretically appealing; it is practically advantageous when data are limited.

MFAFY. On MFAFY, PRISM reaches 73.71% OOD accuracy, compared with 64.86% for ConTeXt-LE and 66.57% for LE-Viz. This dataset is particularly informative because it lacks the rich continuous sensing structure of GLOBEM and

LifeSnaps. Its inputs are mostly categorical self-reports. That PRISM still wins here suggests that the framework's gains do not depend only on sensor-rich inputs. At the same time, some later ablation patterns will show that the relative strength of the spectral stream is smaller here, which is exactly what the architecture's design logic would predict.

The generative path matters. Across all three datasets, $\text{PRISM}_{\text{gen}}$ clearly outperforms $\text{PRISM}_{\text{pred}}$: by 12.85 points on GLOBEM, 12.50 on LifeSnaps, and 8.00 on MFAFY. Because both paths share the same fused representation up to the output heads, this difference is analytically revealing. It indicates that forcing the representation through a generative reasoning-like bottleneck extracts value that a direct MLP classifier cannot fully exploit. The implication is not just that text generation is a prettier output format. It is that generative inference can surface information embedded in the representation more effectively than a one-step discriminative head.

5.12 Distribution-Invariance Compared with LE-Viz

An important lens on PRISM is not only absolute OOD accuracy but the size of the ID to OOD gap. Table 5.4 compares the strongest fine-tuned bimodal VLM with PRISM's generative path.

This table crystallizes one of the central arguments of the chapter. LE-Viz improved OOD performance, but the fine-tuned VLM still exhibited a fairly wide generalization gap. PRISM improves OOD accuracy *and* narrows the gap dramatically, especially on GLOBEM and MFAFY. This pattern is consistent with the claim that frozen backbones and coherence-constrained learning are reducing the system's tendency to over-specialize to the source distribution.

5.13 Why PRISM Works: Evidence from Ablation Studies

The main results show that PRISM works. The ablations clarify *why* it works. They are especially important for a dissertation chapter because they reveal how each design decision contributes to the final behavior.

5.13.1 Dual-path ablation: the two losses regularize each other

Table 5.5 varies only the training objective while holding the architecture fixed. Several conclusions follow directly.

First, the generation path is not a cosmetic auxiliary objective. When the prediction loss is removed, the generative path drops from 79.93% to 71.33% on GLOBEM, from 81.25% to 70.31% on LifeSnaps, and from 73.71% to 68.29% on MFAFY. Thus the prediction loss supplies structure that the generation loss alone does not enforce.

Second, the reverse direction also holds. When the generation loss is removed, the discriminative path worsens from 67.08% to 65.10% on GLOBEM, from 68.75% to 64.06% on LifeSnaps, and from 65.71% to 64.57% on MFAFY. The semantic coherence constraint therefore regularizes the shared representation even for the direct classifier.

Third, learned task weighting is clearly superior to fixed equal weighting. This confirms that the relative usefulness of the two losses is data-dependent and should not be manually imposed. Together these findings justify PRISM's central training claim: the best representation is obtained not by either path alone, but by forcing the model to satisfy both constraints simultaneously.

5.13.2 Learned task weighting reveals the informational advantage of generation

Table 5.6 reports the learned homoscedastic parameters. These values are more than technical details. They tell us how the model itself assesses the relative reliability of the two supervision sources.

The generation loss receives between 2 and 7 times more effective weight than the prediction loss. This makes intuitive sense. A single binary label conveys one bit of direct supervision, whereas a target narrative supplies hundreds of token-level training signals. The generation loss therefore acts as a richer teacher for representation learning. Importantly, the ratio is smaller on MFAFY than on the sensor-rich datasets, which fits the broader evidence that the semantic and spectral advantages are somewhat less dramatic when the data are primarily categorical self-reports.

5.13.3 Modality removal confirms the complementarity claim

Table 5.7 removes each stream in turn and evaluates the generative path.

The ordering is strikingly consistent across datasets. The semantic stream contributes the most, the spectral stream the second most, and the temporal stream the least. This result carries several implications.

First, it validates one of the dissertation's deepest claims: contextual interpretation is not auxiliary decoration on top of behavioral data. It is one of the most powerful sources of transfer. Removing the semantic stream costs 12 to 14 points on every dataset.

Second, the spectral stream matters strongly even though it is derived from the same underlying data instead of a separate sensor type. This supports the idea that

multimodality in LE modeling can arise through complementary representational views of the same source, not only from different external modalities.

Third, the temporal stream contributes least, despite being the only fully trainable non-pretrained backbone. This is powerful evidence for the frozen-prior philosophy. In small-data behavioral settings, pretrained semantic and visual priors adapted lightly at the input side can be more transferable than a model trained entirely from scratch on raw sequences.

Finally, the MFAFY result is particularly informative. Even on a dataset with categorical self-reports, the spectral stream still contributes 9.42 points. This suggests that periodicity and structure in response patterns still matter, though less than on the sensor-rich datasets.

5.14 Interpreting the Final Framework

The empirical evidence supports three broader interpretations of PRISM.

5.14.1 Frozen language coherence acts as a distribution-invariant constraint

The strongest conceptual claim of PRISM is that the frozen LLM defines a coherence surface that is external to the source distribution. When the model is trained jointly under prediction and generation losses, the representation is pushed toward the intersection of discriminative structure and semantic plausibility. Because the LLM cannot adapt, the system cannot solve the task only by teaching the language model the source distribution's shortcuts. Instead, the upstream representation must become reasoning-compatible with a broader prior.

This interpretation is supported indirectly by the ablations. Generation-only training helps but is not enough. Prediction-only training helps but is not enough.

The best results arise when both are active and the generation path remains frozen. In other words, coherence is useful because it is fixed.

5.14.2 Generative inference is not just a reporting interface

$\text{PRISM}_{\text{gen}}$ consistently outperforms $\text{PRISM}_{\text{pred}}$ by a wide margin. This shows that the generative path is not simply providing an explanation after the fact. The act of projecting the fused representation into a soft prompt space and forcing it through a frozen autoregressive decoder serves as an additional reasoning bottleneck. A one-step classifier can separate classes; a narrative generator must articulate a plausible behavioral trajectory. That requirement appears to surface more of the structure encoded in $\mathbf{F}_{\text{fused}}$.

5.14.3 The final framework integrates the whole dissertation

PRISM works by synthesizing and extending the preceding chapters:

- from the ML/DL stage, it retains the importance of temporal measurement evidence;
- from the LLM stage, it retains Meta-Narrative semantic abstraction and prospective narrative generation;
- from the missingness and representation chapter, it inherits the principle that the form of representation determines what the model can learn;
- from LE-Viz, it retains the insight that text alone is structurally incomplete and that visual encodings can preserve information lost in serialization.

The novelty of PRISM lies in how these lessons are constrained together. PRISM is not just "more modalities" or "a bigger model." It is a model that makes the earlier lessons mutually binding.

At the dissertation-question level, this chapter directly addresses **RQ5** and provides strong support for **H5**. Together with Chapter 4, it also reinforces the evidence for **RQ4**/**H4** by showing that multimodal structure preservation is strongest when it is paired with explicit training constraints.

5.15 Limitations and Scope of the Final Framework

Although PRISM is the strongest framework in the dissertation, it is not without limits.

First, the system still depends on carefully designed representational views. The temporal, spectral, and semantic streams were not discovered automatically; they were engineered based on the structure of LE data. Poorly chosen views would weaken the entire framework.

Second, the spectral stream currently relies on a fixed Morlet wavelet design. Alternative wavelet families or adaptive spectral parameterization may further improve sensitivity to specific behavioral regimes [85].

Third, while freezing promotes transfer and efficiency, it also constrains the ceiling of domain adaptation. There may be settings where selective unfreezing of certain layers improves performance, though at a likely cost to invariance.

Fourth, the chapter's strongest evidence is concentrated on three completed datasets. A TILES-style occupational validation setting was prepared in the PRISM project materials but not fully finalized with stable results at the time of this dissertation chapter. Rather than inserting incomplete results, this chapter focuses

on the three completed evaluations where the evidence is strong and internally consistent.

These limitations do not diminish the role of PRISM. Rather, they clarify what kind of answer it provides. PRISM is not the final word on LE modeling in general. It is the strongest answer this dissertation offers to the specific problem of building generalizable and reliable LE representations under severe distribution shift.

5.16 Chapter Summary

This chapter presented PRISM as the final technical framework of the dissertation. The chapter argued that the main obstacle in LE forecasting is not simply insufficient model capacity, but insufficient representational constraint. PRISM addressed this by decomposing each trajectory into temporal, spectral, and semantic views, fusing them through directed cross-modal attention and instance-specific gates, and training the resulting representation under a dual objective that combines direct prediction with frozen-language narrative coherence.

Empirically, PRISM set the strongest OOD results in the dissertation: 79.93% on GLOBEM, 81.25% on LifeSnaps, and 73.71% on MFAFY. It improved over the strongest earlier text-only and multimodal baselines while also narrowing the ID to OOD gap relative to the fine-tuned VLM stage. The ablations showed that neither prediction nor generation is sufficient alone, that learned task weighting consistently favors the richer generation signal, and that the semantic and spectral streams contribute strongly across diverse datasets.

At the level of the full dissertation arc, PRISM realizes the dissertation's central claim in its most complete form: reliable cross-distribution modeling of LE data requires not just stronger models, but representations that jointly

preserve behavioral evidence, temporal dynamics, and contextual meaning under learning constraints that are hard to satisfy through source-specific shortcuts. With this chapter, the technical arc of the dissertation is complete. The next chapter synthesizes the contributions of the dissertation as a whole, revisits the research questions, and outlines the broader implications and future directions of reliable multimodal AI for LE data.

Table 5.3: Cross-distribution generalization results. Bold marks the best OOD result for each dataset. The shaded rows correspond to PRISM's two inference paths.

Dataset	Method	In-Distribution (ID)				Out-of-Distribution (OOD)			
		Acc	P	R	F1	Acc	P	R	F1
GLOBEM	*Time-Series Models*								
	PatchTST	53.58	53.73	53.58	53.01	49.88	49.24	49.88	49.16
	iTransformer	54.61	54.61	54.61	54.61	51.06	51.07	51.06	51.07
	LLM Models								
	LLM-Complete	69.96	71.56	68.42	69.96	65.94	67.95	68.52	68.23
	LLM-Summary	69.51	72.22	67.24	69.65	62.43	65.97	63.57	64.75
	LLM-Clustered	70.05	71.30	69.37	70.32	66.44	67.92	69.09	68.50
	ConTeXt-LE	73.99	75.93	71.93	73.87	67.40	68.81	70.00	69.40
	VLM Models								
	Time-VLM	57.34	55.92	56.48	56.20	51.43	49.87	50.21	50.04
	Bimodal VLM	79.37	83.33	80.77	82.03	72.86	75.58	74.28	74.92
	$\text{PRISM}_{\text{pred}}$	70.85	76.39	70.10	68.74	67.08	73.49	65.40	63.22
	$\text{PRISM}_{\text{gen}}$	81.17	83.39	80.75	80.69	**79.93**	**82.53**	**79.06**	**79.14**
LifeSnaps	*Time-Series Models*								
	PatchTST	47.83	47.83	47.83	47.83	43.75	41.23	43.75	40.47
	iTransformer	52.17	46.81	52.17	48.83	48.44	48.14	48.44	47.99
	LLM Models								
	LLM-Complete	58.82	77.78	58.33	66.67	54.84	50.00	57.14	53.33
	LLM-Summary	47.06	40.00	57.14	47.06	46.88	36.67	42.31	39.29
	LLM-Clustered	70.59	80.00	72.72	76.19	62.50	52.94	69.23	60.00
	ConTeXt-LE	64.71	77.78	63.64	70.00	67.19	63.89	74.19	68.66
	VLM Models								
	Time-VLM	49.28	47.11	47.86	47.48	47.36	45.02	45.71	45.36
	Bimodal VLM	82.35	85.71	75.00	80.00	71.88	83.33	71.43	76.92
	$\text{PRISM}_{\text{pred}}$	75.00	85.71	66.67	66.67	68.75	74.97	67.94	66.07
	$\text{PRISM}_{\text{gen}}$	87.50	87.50	90.00	87.30	**81.25**	**83.46**	**80.84**	**80.78**
MFAFY	*Time-Series Models*								
	PatchTST	67.39	63.66	67.39	64.34	50.57	49.68	50.57	44.31
	iTransformer	53.26	51.77	53.26	52.47	44.29	43.05	44.29	42.42
	LLM Models								
	LLM-Complete	60.66	56.67	60.71	58.62	57.14	50.55	60.53	55.09
	LLM-Summary	57.38	48.28	56.00	51.85	53.43	52.02	52.94	52.48
	LLM-Clustered	63.93	62.96	69.37	60.71	62.86	57.47	64.10	60.61
	ConTeXt-LE	70.49	65.22	60.00	62.50	64.86	61.11	67.48	64.14
	VLM Models								
	Time-VLM	53.41	51.76	52.33	52.04	48.27	46.14	46.82	46.48
	Bimodal VLM	77.05	75.76	80.65	78.12	66.57	64.47	60.87	62.62
	$\text{PRISM}_{\text{pred}}$	69.81	67.74	65.91	66.35	65.71	62.82	59.63	59.40
	$\text{PRISM}_{\text{gen}}$	75.47	74.18	72.42	73.01	**73.71**	**78.03**	**66.50**	**66.77**

Table 5.4: ID to OOD accuracy gap. Smaller values indicate more distribution-invariant behavior.

Method	**GLOBEM**	**LifeSnaps**	**MFAFY**
Bimodal VLM	6.51	10.47	10.48
$\text{PRISM}_{\text{gen}}$	**1.24**	**6.25**	**1.76**

Table 5.5: Dual-path ablation. Pred and Gen denote the prediction and generative inference paths respectively. Gray values indicate output heads that received no direct training signal under that configuration.

		GLOBEM		**LifeSnaps**		**MFAFY**	
	Objective	Pred	Gen	Pred	Gen	Pred	Gen
(a)	$\mathcal{L}_{\text{pred}}$ only	65.10	59.12	64.06	64.06	64.57	55.71
(b)	$\mathcal{L}_{\text{gen}}$ only	55.86	71.33	64.06	70.31	52.29	68.29
(c)	Fixed equal weighting	68.26	75.73	70.31	76.56	67.43	71.71
(d)	Learned homoscedastic weighting	67.08	**79.93**	68.75	**81.25**	65.71	**73.71**

Table 5.6: Learned homoscedastic weighting parameters at convergence. Higher effective weight indicates greater influence on training.

	Learned s		**Effective weight** e^{-s}		**Ratio**
Dataset	s_1 Pred	s_2 Gen	Pred	Gen	Gen/Pred
GLOBEM	1.44	-0.39	0.24	1.48	6.2×
LifeSnaps	0.76	-1.17	0.47	3.22	6.9×
MFAFY	0.95	0.28	0.39	0.76	2.0×

Table 5.7: Modality removal ablation on the generative path. Δ denotes the OOD accuracy drop from the full system.

	GLOBEM		**LifeSnaps**		**MFAFY**	
Configuration	Acc	Δ	Acc	Δ	Acc	Δ
Full system	79.93	—	81.25	—	73.71	—
Without text (vision + temporal)	65.62	-14.31	68.75	-12.50	61.14	-12.57
Without vision (text + temporal)	68.81	-11.12	71.88	-9.37	64.29	-9.42
Without temporal (text + vision)	71.13	-8.80	73.44	-7.81	67.71	-6.00

Table 5.8: How PRISM synthesizes the dissertation arc.

Stage in dissertation	**Primary representational advance**	**Main limitation left unresolved**	**How PRISM incorporates or answers it**
Traditional ML/DL	Direct modeling of temporal measurements	Weak contextual reasoning and poor shift robustness	Retains measurement evidence as the temporal stream but does not rely on it alone
Narrative LLM modeling	Semantic contextualization and generative forecasting	Text serialization weakens native structure	Retains meta-narratives as the semantic stream and the generative forecasting paradigm
Missingness / context stage	Representation itself governs what can be learned	Single-view representations remain fragile under shift	Treats representation as multi-view and complementary instead of singular
LE-Viz multimodal modeling	Visual preservation of structure lost in text-only inputs	Fine-tuned multimodal systems can still overfit or suppress modalities	Retains multimodality but constrains it through frozen backbones and directed fusion
PRISM	Frozen multimodal constraints plus semantic coherence regularization	—	Final integrated answer of the dissertation

Chapter 6

Conclusion and Future Directions

6.1 Introduction

This dissertation began with a simple but consequential observation: human-centered longitudinal data is difficult to model not because it is merely large, noisy, or incomplete, but because it is *contextual*, *heterogeneous*, *partially observed*, and *distributionally unstable*. These properties make LE data fundamentally different from the structured numerical datasets for which many standard machine learning pipelines were originally designed. In LE settings, the same behavioral pattern can mean different things depending on context, the absence of a response may itself carry signal, and the distributions encountered at deployment may differ substantially from those seen during training [4–7, 16, 61, 62].

The central question of the dissertation has therefore been the following:

How can one build generalizable and reliable models of LE data?

The answer developed across the dissertation is that reliable LE modeling requires more than larger models, more parameters, or stronger optimization. It requires a different way of thinking about *representation*. Throughout the dissertation, the core argument has been that no single representation is sufficient. Numerical

encoding preserves measurement fidelity but weakens semantic interpretation. Textual encoding preserves context and behavioral meaning but can collapse structural relationships into a one-dimensional token stream. Visual encoding preserves spatial and temporal organization but can underuse semantic context if applied alone. A robust solution therefore requires complementary views of the same trajectory, together with learning objectives and architectural constraints that discourage source-specific shortcuts and support transfer under distribution shift [1–3, 20–22].

This final chapter synthesizes the dissertation's contributions and explains how the research program developed from traditional ML and DL forecasting, to contextual language modeling, to missingness-aware representation, to narrative-based cross-distribution generalization, to multimodal visuospatial modeling, and finally to the frozen-backbone multimodal PRISM framework. It also identifies the broader methodological lessons of the work, discusses its limitations, and outlines future directions for reliable AI on LE data [3, 21, 22].

6.2 Central Claim Revisited

The central claim of this dissertation can be restated as follows:

> **Generalizable and reliable modeling of LE data requires multimodal representations that jointly preserve semantic context, temporal structure, and behavioral dynamics, together with learning objectives that constrain the model toward distribution-invariant reasoning rather than distribution-specific shortcuts.**

This claim was established progressively across the dissertation. The early stages of the work focused on forecasting accurately in small educational datasets.

Later stages reframed the harder problem as *generalization under change*: models needed to remain dependable across new years, new cohorts, new institutions, and new behavioral regimes. As this progression unfolded, the central bottleneck became clear. The key challenge was alignment between the structure of LE data and the representations used to model it.

That insight changed the direction of the dissertation. Instead of asking only which model class performs best, the dissertation increasingly asked: what must be preserved about an LE trajectory so that a model can reason about it robustly? The answer, in retrospect, required three linked commitments.

First, LE data must be interpreted *contextually*. Raw values are not enough because the meaning of a behavioral pattern often depends on surrounding circumstances. Second, LE data must be represented *structurally*. A trajectory is not a bag of independent variables; it is a temporally organized object with relationships across features and across time. Third, LE models must be trained *under constraint*. If the learned representation is allowed to drift too freely toward the source distribution, strong ID results may coexist with brittle OOD behavior. The strongest results in this dissertation emerged only when these three commitments were pursued jointly [3, 21, 22].

6.3 From a Forecasting Problem to a Representation Problem

One of the most important intellectual outcomes of the dissertation is that LE forecasting should no longer be viewed narrowly as a sequence prediction problem. It is more accurately understood as a *representation problem under distribution shift*. This reframing matters because it explains why multiple seemingly different technical contributions in the dissertation are in fact parts of one coherent story.

Traditional numerical pipelines begin with a representation that is already lossy for LE data. They flatten trajectories, aggregate away local temporal variation, and treat features as if their meanings are fixed across settings. In contrast, the language-based stages of the dissertation showed that verbalization and contextualization can reveal structure that is invisible to purely numerical models. The missingness-focused stage showed that even the *absence* of information can be part of the representational object. ConText-LE then showed that better abstraction and better output formulation together can materially improve cross-distribution performance. LE-Viz showed that text alone is not enough because one-dimensional serialization suppresses important Feature $\times$ Time structure. PRISM then demonstrated that even multimodality is not sufficient unless the model is prevented from drifting toward brittle task-specific shortcuts.

Across the dissertation, the technical trajectory is also a conceptual trajectory. It moves from asking *how to predict*, to asking *how to represent*, and finally to asking *how to constrain* representation learning so that prediction remains reliable when the data distribution changes. This is the deepest unifying idea of the dissertation.

6.4 Summary of the Dissertation's Technical Arc

Each stage of the dissertation solved a problem exposed by the previous stage and contributed evidence for the research questions and working hypotheses introduced in Chapter 1. Stage 1 addresses **RQ1** and establishes the limitations that motivate **H1**. Stages 2 and 4 address **RQ2** and support **H1** and **H3** through contextual language modeling and narrative forecasting. Stage 3 addresses **RQ3** and supports **H2** through missingness-aware semantic representation. Stage 5 addresses **RQ4** and supports **H4** through structure-preserving multimodal mod-

eling. Stage 6 addresses **RQ5** and supports **H5** through frozen-prior constrained multimodal learning.

6.4.1 Stage 1: Traditional ML and DL as necessary but limited baselines

The dissertation first established the role of classical machine learning and deep learning methods as important reference points for LE forecasting. Models such as support vector machines, random forests, recurrent networks, and modern time-series architectures provide useful baselines because they are efficient, familiar, and often competitive in simpler structured settings [8–10, 70, 71].

However, the dissertation also showed that these methods are fundamentally limited for the full LE problem. They typically assume that features have fixed meanings, that aggregation is mostly harmless, and that the learning objective can be satisfied by discriminative separation over flattened or weakly structured inputs. Those assumptions become increasingly fragile as data becomes more subjective, sparse, missingness-rich, and context-dependent. The GLOBEM benchmark made that difficulty especially clear by showing that conventional approaches can fall near chance under realistic cross-distribution evaluation [16]. The lesson of this stage was therefore not that classical methods are useless, but that they reveal the structural difficulty of the problem and provide the contrast needed to justify richer representation strategies.

6.4.2 Stage 2: Contextual language modeling as a semantic shift in representation

The next stage of the dissertation introduced LLMs as a new representational paradigm for LE data. Instead of treating trajectories only as numeric arrays, the work reformulated them as semantically interpretable sequences that could be

verbalized, contextualized, and forecasted through language generation [20, 26, 27].

This move mattered for two reasons. First, language models brought strong pre-trained priors for contextual reasoning, which are especially useful when data is limited but semantically rich. Second, verbalization provided a more natural interface for combining distal context, proximal behavioral patterns, and interpretable forecasting targets. The dissertation's early educational forecasting work showed that this approach could outperform traditional baselines in small-data settings not because language models are universally stronger, but because the data representation better matched the reasoning capabilities of the model [26, 27].

This stage established a durable principle that remained central throughout the rest of the dissertation: performance gains in LE modeling do not arise only from stronger architectures; they often arise from aligning the representation with what the model class is good at interpreting.

6.4.3 Stage 3: Missingness-aware representation as semantic modeling of absence

As the dissertation moved toward richer LE settings, missingness emerged as a central difficulty rather than a secondary preprocessing nuisance. In many human-centered datasets, nonresponse is not random. A missing self-report, skipped prompt, or absent measurement may correlate with mood, stress, disengagement, or contextual burden. Classical missing-value methods often treat these cases as entries to repair numerically. The dissertation instead argued that missingness in LE data is often part of the behavioral signal itself [1, 31, 32, 61, 62].

This idea led to a missingness-aware representational framework in which missing values were encoded through contextually meaningful descriptors rather than only replaced by numeric surrogates. In CRILM, LLMs generated semantically

grounded textual descriptors for missing entries, allowing smaller downstream language models to reason over incomplete data in a more context-aware way [1]. In the three-tier engagement forecasting framework, this logic became part of a broader pipeline that combined LLM-informed imputation, feature selection, and downstream LE forecasting from qualitative trajectories [2].

The lesson of this stage was foundational for the dissertation. Missingness should not be treated only as a defect in the data matrix. In many LE problems, it is also a clue about the trajectory being modeled. This was one of the most important expansions of the dissertation's original forecasting viewpoint.

6.4.4 Stage 4: ConText-LE and narrative-based cross-distribution generalization

Once contextualized textual representation had proven useful, the dissertation turned directly to the challenge of cross-distribution generalization. Earlier LLM-based LE systems improved forecasting, but they did not yet systematically optimize representation and output formulation for transfer across time periods, cohorts, and contexts. ConText-LE was developed to address that gap [3].

Its two most important ideas were *Meta-Narrative* representation and *Prospective Narrative Generation*. Meta-Narratives transformed a raw LE window into a semantically rich behavioral account that emphasized salient trends, interactions, and contextual meaning. Prospective Narrative Generation reframed prediction as future-oriented narrative generation rather than direct binary classification. This change was deeper than a cosmetic switch in output format. It aligned the training objective more closely with how language models express contextual reasoning.

The empirical gains were substantial. On GLOBEM, ConText-LE reached 67.40% OOD accuracy and 69.40% OOD F1, far beyond the weak cross-distribution behavior of conventional time-series baselines [3]. On LifeSnaps and MFAFY, the

same narrative formulation also produced the strongest text-only cross-distribution results in the dissertation, reaching 67.19% and 64.86% OOD accuracy respectively [3]. This stage therefore established that representation and output formulation are both first-class determinants of generalization.

At the same time, ConText-LE exposed a remaining limitation. Even the strongest narrative abstraction is still a one-dimensional textualization of a richer object. It improves semantic reasoning, but it cannot fully preserve the native structural organization of LE data.

6.4.5 Stage 5: LE-Viz and structure-preserving multimodal modeling

The next stage responded directly to that limitation. If text alone collapses structure, then another modality is needed to preserve what serialization discards. LE-Viz developed that idea by pairing textual Meta-Narratives with visual encodings such as line charts and heatmap chains, allowing a VLM to process both semantic interpretation and structure-preserving spatial organization [19,21,67,68].

This was a decisive step because it changed the role of multimodality in the dissertation. Multimodality was not introduced as an embellishment or as a way to use a larger model. It was introduced as a *representational correction*. The visual channel restored aspects of temporal locality, adjacency, and cross-feature organization that the text channel could not preserve equally well.

The results confirmed that this correction mattered. On GLOBEM, LE-Viz with chart-based encoding reached 72.86% OOD accuracy and 74.92% OOD F1, surpassing the best text-only ConText-LE configuration by a meaningful margin [21]. On LifeSnaps and MFAFY, LE-Viz likewise improved over text-only baselines, although the gains varied by dataset and were strongest where the visual channel had naturally continuous temporal structure to exploit [21]. This stage showed

that multimodality is not optional for LE data when the goal is to preserve the structure of the underlying trajectory.

Yet LE-Viz also exposed another challenge. End-to-end multimodal adaptation can be computationally heavy, may become modality-imbalanced, and still leaves a nontrivial ID to OOD gap. That problem motivated the final stage of the dissertation.

6.4.6 Stage 6: PRISM and frozen multimodal constraints for reliable transfer

The final technical stage culminated in PRISM, which integrated the key lessons of the dissertation into a frozen-backbone multimodal framework [22]. PRISM decomposed each LE trajectory into three complementary views: a temporal measurement stream, a spectral dynamics stream, and a semantic interpretation stream. These streams were then fused through directed cross-modal interaction while the major pretrained backbones remained frozen.

The most important idea in PRISM was that reliability depends not only on adding complementary views, but on *constraining how those views are learned*. A prediction loss alone can often be satisfied through source-specific shortcut learning. PRISM therefore introduced a dual-objective logic in which the representation had to support both direct prediction and coherent narrative generation under a frozen language prior. In this formulation, the frozen language model acts as a semantic anchor rather than as a freely adapting decoder [22].

This proved to be the strongest answer in the dissertation. PRISM reached 79.93% OOD accuracy on GLOBEM, 81.25% on LifeSnaps, and 73.71% on MFAFY [22]. Just as importantly, it narrowed the ID-OOD gap relative to the fine-tuned LE-Viz stage, especially on GLOBEM and MFAFY. The result was not merely a stronger model, but a stronger formulation of what reliable LE modeling requires.

Table 6.1: The connected technical arc of the dissertation. Each stage solved a limitation exposed by the previous one and introduced a more faithful or more constrained LE representation.

Stage	Primary representation	Problem addressed	What it established	What remained unresolved
Traditional ML/DL	Flattened or weakly structured numeric trajectories	Baseline LE forecasting	LE forecasting is harder than standard supervised prediction; purely numeric modeling is brittle under shift	Weak contextual reasoning, limited missingness handling, and poor OOD robustness
Contextual language modeling	Verbalized and contextualized LE narratives	Aligning inputs with LM priors	LLMs can outperform numeric baselines when representation matches contextual reasoning strengths	Textualization still compresses structure and cannot fully preserve trajectory geometry
Missingness-aware modeling	Descriptor-based incomplete-data representation	Modeling missingness as signal	Missingness can be semantically informative and should be represented explicitly	Better missingness handling alone does not solve OOD generalization
ConText-LE	Meta-Narratives plus Prospective Narrative Generation	Cross-distribution text-only forecasting	Input abstraction and output formulation jointly matter for transfer	One-dimensional text remains structurally lossy
LE-Viz	Visual-text multimodal encoding	Recovering Feature $\times$ Time structure lost in text serialization	Multimodality preserves complementary structure and improves OOD performance	End-to-end multimodal adaptation remains costly and can still overfit source-specific patterns
PRISM	Frozen temporal, spectral, and semantic streams under dual objectives	Reliable multimodal generalization under shift	Constrained fusion with frozen priors yields the strongest and most stable transfer behavior in the dissertation	Open questions remain on uncertainty, intervention, and scaling to broader domains

6.5 Summary of Main Contributions

Taken together, the dissertation makes the following principal contributions.

6.5.1 Contribution 1: A representation-first view of LE forecasting

The dissertation shows that LE forecasting should not be treated merely as a numerical time-series prediction problem. It should be treated as a problem of representing behavioral trajectories in ways that preserve context, temporal organization, and meaning under distribution shift [3, 20–22].

6.5.2 Contribution 2: Contextual language modeling for small-data LE settings

The dissertation demonstrates that verbalization, personalization, and contextualization make pretrained language models effective tools for LE forecasting in small-data educational settings. This establishes language-based transfer learning as a viable alternative to conventional numeric modeling when the target signals are semantically rich and behaviorally contextual [20, 26, 27].

6.5.3 Contribution 3: Missingness-aware representation

The dissertation contributes a context-aware treatment of missingness in which missing values are represented through semantically meaningful descriptors rather than reduced only to numeric surrogates. This makes incomplete data part of the modeling signal rather than merely a preprocessing inconvenience [1, 2].

6.5.4 Contribution 4: Narrative-based cross-distribution generalization

The dissertation introduces narrative-based LE forecasting strategies that improve OOD generalization by aligning both the input representation and the output formulation with the reasoning strengths of language models. Meta-Narrative representation and Prospective Narrative Generation together establish the strongest text-only generalization stage in the dissertation [3].

6.5.5 Contribution 5: Structure-preserving multimodal LE modeling

The dissertation introduces visual-text LE representation, showing that visuospatial encoding can recover trajectory structure that text-only modeling weakens. This establishes the importance of multimodal representation in human-centered

longitudinal forecasting and clarifies when the visual channel is especially beneficial [21].

6.5.6 Contribution 6: Frozen multimodal constraints for reliable transfer

Finally, the dissertation contributes a frozen-backbone multimodal framework that integrates temporal, spectral, and semantic trajectory views under both prediction and coherence objectives. This provides the strongest technical answer in the dissertation to the problem of generalizable and reliable LE modeling [22].

6.6 Empirical Synthesis Across the Dissertation

The empirical results form a consistent pattern that mirrors the core argument. Each time the representation became more faithful to the structure of LE data, and each time the learning setup became more disciplined with respect to generalization, OOD behavior improved.

This pattern is most visible in the cross-distribution benchmarks that recur across the later chapters. Table 6.2 summarizes representative OOD milestones from the strongest text-only, multimodal, and constrained-multimodal stages of the dissertation.

Table 6.2: Representative OOD accuracy milestones across the dissertation. For GLOBEM, the older baseline is the traditional benchmark reported in prior work; for LifeSnaps and MFAFY, it is the strongest time-series baseline used in the dissertation comparisons.

Dataset	Older Baseline	ConText-LE	LE-Viz	PRISM
GLOBEM	52.80	67.40	72.86	**79.93**
LifeSnaps	48.44	67.19	71.88	**81.25**
MFAFY	50.57	64.86	66.57	**73.71**

Several points follow from this progression.

First, the gains are *stepwise and theory-consistent*. ConText-LE improves over conventional baselines because narrative abstraction and generative formulation better match how language models reason about context. LE-Viz then improves further because it restores structural information that text alone suppresses. PRISM improves again because multimodality is no longer left unconstrained; instead, it is organized through frozen semantic priors and disciplined fusion [3, 21, 22].

Second, the pattern is *dataset-sensitive in an interpretable way*. The visual and spectral channels help most on sensor-rich datasets such as GLOBEM and LifeSnaps, where temporal dynamics have strong shape and continuity. On MFAFY, gains remain meaningful but are smaller at the LE-Viz stage because the raw inputs are more categorical and qualitative. This boundary case is important because it suggests that the dissertation's frameworks are not succeeding through accidental overfitting to one data type. Instead, they behave differently when the representational affordances of the modality change, which is exactly what a faithful theory would predict [20–22].

Third, the final stage does not only improve OOD accuracy; it also narrows the ID-OOD gap. That is a particularly important signal because it suggests that the final model is not merely optimizing harder on the source distribution. It is learning representations that are more stable when the evaluation regime changes. In this sense, the strongest evidence of the dissertation is not any single number, but the fact that better OOD performance repeatedly coincides with representational designs that are more semantically grounded, more structurally faithful, and more strongly constrained [21–23].

6.7 Design Principles Emerging from the Dissertation

Beyond its specific models, the dissertation contributes a set of design principles for future LE modeling systems. These principles are among the most durable outcomes of the research because they are not tied to a single architecture or dataset.

6.7.1 Principle 1: Representation matters as much as model scale

One of the clearest lessons of the dissertation is that stronger results did not come simply from moving to larger models. Rather, they came from changing how the data was represented: from raw numeric sequences, to contextualized language, to missingness-aware narratives, to multimodal visual-text views, and finally to constrained multimodal fusion. In human-centered domains, representational design can be at least as important as raw parameter count [3, 21, 22, 59].

6.7.2 Principle 2: Context should be made explicit, not assumed to be recoverable

A major reason LLM-based approaches helped is that they made contextual meaning explicit in the representation. This suggests a broader principle: if the meaning of a trajectory depends on surrounding circumstances, then context should be encoded deliberately rather than left implicit in raw values. Future systems that ignore this principle may continue to mistake context-sensitive behavior for context-free signal [3, 26, 27].

6.7.3 Principle 3: Missingness is often structure, not only corruption

The missingness work in the dissertation suggests that incomplete-data handling in LE settings should be rethought. In many human-centered problems, absence is not a nuisance alone. It can reflect disengagement, overload, avoidance, or other meaningful behavioral states. Systems that erase this information through purely mechanical imputation may lose signal that matters for reliable forecasting [1, 2, 61, 62].

6.7.4 Principle 4: Output formulation affects generalization

ConText-LE showed that how a model is asked to express its prediction matters for transfer. Prospective Narrative Generation was not merely a more verbose output interface; it changed the learning problem in a way that better matched the strengths of language models. This suggests that output design should be regarded as part of the modeling problem rather than as an afterthought [3].

6.7.5 Principle 5: Multimodality should preserve complementary structure

The multimodal stages of the dissertation were strongest when each modality preserved something the others lost. In this work, text preserved semantic context, visual encoding preserved trajectory structure, and spectral transformation preserved dynamic variation. This is a more disciplined view of multimodality than simply concatenating more inputs. It suggests that effective multimodal LE systems should be designed around complementary informational roles [19, 21, 22].

6.7.6 Principle 6: Freezing can be a generalization mechanism, not only an efficiency trick

A final principle emerging from PRISM is that frozen backbones can improve transfer not merely by saving compute, but by preventing semantic drift toward the source distribution. In other words, freezing can act as a regularization mechanism that protects the model from over-specializing to the training environment. This is a broader insight that may matter well beyond LE data [22].

6.8 Broader Implications

The contributions of this dissertation extend beyond the specific datasets and tasks studied here. More broadly, the work suggests several implications for AI in human-centered domains.

6.8.1 Implications for education

In educational settings, the dissertation suggests that student forecasting should move beyond narrow performance histories and toward richer longitudinal views that include behavioral, affective, and non-cognitive context. It also suggests that the strongest forecasting systems may be those that preserve interpretability and contextual meaning rather than those that optimize only over compact numerical summaries [2, 8, 9, 20].

6.8.2 Implications for mental health and behavioral medicine

In behavioral health contexts, the work reinforces the promise of LE data for understanding dynamic risk, but it also highlights the dangers of distributional brittleness. Models that appear effective within one cohort or time period may

fail badly under new conditions. The dissertation therefore supports a shift toward generalization-first evaluation in digital phenotyping and personal sensing applications [3, 7, 16, 22, 28].

6.8.3 Implications for foundation model adaptation

The dissertation also contributes to a broader conversation in foundation-model research. Much current work assumes that stronger adaptation or broader fine-tuning is the path to better performance. This dissertation suggests a more careful alternative: in small, high-stakes, distributionally unstable settings, performance may improve more by constraining adaptation and redesigning representation than by simply making the optimization loop more flexible [22, 39, 58, 59].

6.8.4 Implications for reliable AI

Finally, the dissertation contributes to a more general point about reliable AI. In human-centered domains, reliability cannot be reduced to average accuracy on a held-out split. Reliability also depends on whether the representation remains meaningful under change, whether the model avoids overconfident shortcut behavior, and whether the system preserves enough structure to support interpretation and future intervention. The methods developed here do not solve all aspects of reliability, but they clarify an important part of the path forward [3, 22, 23].

6.9 Limitations of the Dissertation

Although the dissertation advances the state of LE modeling in several important ways, it also has limitations that should be acknowledged clearly.

6.9.1 Dataset scope

Most of the empirical development is centered on educational and behavioral datasets. These settings are rich, socially important, and methodologically demanding, but they do not exhaust the space of possible LE domains. Additional validation in healthcare, workplace analytics, digital therapeutics, or longitudinal clinical monitoring would strengthen the generality of the framework [7, 16, 28].

6.9.2 Dependence on careful representation engineering

Many of the gains in this dissertation depend on thoughtful design of narrative prompts, descriptor styles, visual encodings, and representational decompositions. This is a strength in one sense because it reveals how important representation is. But it is also a limitation because such design choices may require domain expertise and careful iteration. More automated representation discovery remains an open problem [3, 21, 22].

6.9.3 Reliance on strong pretrained external models

Several stages of the dissertation rely on large pretrained language or VLMs for representation generation, narrative synthesis, or multimodal inference. Although this is part of the dissertation's methodological point, it does mean that some parts of the pipeline depend on external foundation-model capabilities that may change over time or differ across providers [39, 58, 59].

6.9.4 Limited treatment of uncertainty, fairness, and causality

The dissertation focuses primarily on representation, generalization, and multimodality. It does not fully resolve calibrated uncertainty quantification, fairness

across subpopulations, privacy-preserving deployment, or causal interpretability. These are especially important issues in human-centered AI, and future work will need to engage them directly rather than assume that stronger prediction alone is sufficient [6, 7, 23].

6.9.5 Forecasting rather than intervention

The models developed here are forecasting systems. They are designed to anticipate future states, not yet to recommend or optimize interventions. In many real applications, however, the ultimate goal is not only to predict who may struggle next week, but to decide what kind of support might change that trajectory. That transition from forecasting to action remains largely beyond the current dissertation.

6.10 Future Directions

The research program developed in this dissertation opens several promising next directions.

6.10.1 Uncertainty-aware LE modeling

A major next step is to incorporate calibrated uncertainty into LE forecasting. In high-stakes human-centered settings, a model should not only make accurate forecasts; it should also indicate when its predictions are uncertain or distributionally fragile. Future work could integrate conformal methods, ensemble strategies, or other calibration-aware techniques into the multimodal pipelines developed here, especially under OOD evaluation [23].

6.10.2 Intervention-aware and decision-support modeling

A second direction is to move from passive forecasting toward decision support. Instead of predicting only what is likely to happen, future systems should help reason about how a trajectory may change under intervention, support, or environmental disruption. This would connect LE modeling more directly to educational advising, mental health triage, and personalized assistance systems.

6.10.3 Causal and counterfactual trajectory reasoning

Relatedly, future systems should move beyond predictive robustness toward causal reasoning. Many patterns in LE data are predictive without being causally actionable. A key next step is to distinguish between features that correlate with future outcomes and features that reflect mechanisms whose alteration could plausibly change those outcomes. Counterfactual or intervention-sensitive LE modeling would be especially valuable in settings where the purpose of prediction is to guide support rather than merely to classify risk.

6.10.4 Richer multimodal LE ecosystems

The multimodal frameworks in this dissertation focus primarily on streams derived from the same underlying trajectory: temporal measurements, semantic narratives, spectral dynamics, and visual encodings. Future work could incorporate additional modalities such as free-text reflections, audio, video, social interaction graphs, geospatial traces, intervention histories, and environmental context. This would create a richer and more realistic testbed for multimodal human-centered AI [19, 21, 22].

6.10.5 Foundation models specialized for LE data

Much of the dissertation relies on adapting general-purpose language and VLMs. A natural long-term direction is the development of foundation models pretrained specifically for LE data. Such models could learn priors over trajectory dynamics, missingness patterns, and contextual behavioral meaning in ways that general-purpose models are not explicitly trained to do [39, 58, 59].

6.10.6 Automated representation discovery

The dissertation has shown repeatedly that representation design is central. A long-term research goal is therefore to build systems that can discover or adapt their own best representational decomposition for a given LE task, rather than relying entirely on handcrafted narratives, descriptor templates, visual encodings, or spectral transforms. Success in this direction would make the dissertation's representational insights more scalable across domains [3, 21, 22].

6.10.7 Continual, online, and privacy-preserving LE modeling

Real-world LE systems are unlikely to operate in a single static training-deployment cycle. They will need to handle evolving cohorts, new sensing regimes, changing institutional contexts, and strong privacy expectations. Future work should therefore investigate continual adaptation, online generalization monitoring, and privacy-preserving or federated variants of LE modeling that remain faithful to the reliability goals developed in this dissertation.

6.11 Concluding Perspective

At its core, this dissertation has been about building AI systems that do not merely fit data, but *understand trajectories more reliably*. LE data demands this because it reflects people, not just measurements. It records patterns of behavior, uncertainty, absence, change, and context that do not submit easily to flat tabular assumptions or purely numerical forecasting [20, 22].

The dissertation's answer has therefore been progressive and cumulative. It began by asking whether traditional numerical methods were enough. They were not. It then showed that contextual language modeling was a major step forward. It demonstrated that missingness itself had to be represented meaningfully. It then showed that better narrative abstraction and generative formulation improved cross-distribution robustness. It established that multimodality was necessary to preserve structure that text alone lost. Finally, it showed that the strongest answer combined multimodality with frozen semantic constraints so that the system remained anchored to transferable behavioral abstractions [1, 3, 21, 22, 26].

The final result is a coherent research argument:

> *Generalizable and reliable LE modeling is fundamentally a problem of disciplined representation.*

That statement is the unifying conclusion of the dissertation. It explains why traditional numeric models underperform, why contextual language models help, why missingness had to be reinterpreted, why narrative generation improved transfer, why visual encoding mattered, and why the final framework depended on frozen multimodal constraints. The argument is methodological, empirical, and conceptual at once.

Bibliography

[1] A. Hayat and M. R. Hasan, "A context-aware approach for enhancing data imputation with pre-trained language models," in *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, 2025, pp. 5668–5685. (document), 1.1, 1.2.4, 1.6, 1.10.3, 1.13, 1.14.3, 2.5.4, 2.7.3, 3.9.5, 3.6, 3.9.6, 3.7, 3.9.6, 3.4, 3.5, 3.6, 3.7, 3.9.7, 3.8, 3.11, 6.1, 6.4.3, 6.5.3, 6.7.3, 6.11, A.3, A.6, A.8, A.7, B.1, B.4

[2] A. Hayat, H. Martinez, B. Khan, and M. R. Hasan, "A three-tier llm framework for forecasting student engagement from qualitative longitudinal data," in *Proceedings of the 29th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2025, pp. 334–347. (document), 1.1, 1.4.3, 1.6, 1.14.4, 2.1, 2.4.3, 2.5.1, 2.5.2, 2.5.4, 2.6.3, 2.2, 2.7.3, 3.1, 3.1, 3.4, 3.3, 3.9, 3.9.1, 3.9.9, 3.9, 3.8, 3.9, 3.9.10, 3.10, 3.9.10, 3.11, 6.1, 6.4.3, 6.5.3, 6.7.3, 6.8.1, B.1, B.4

[3] A. Hayat, B. Khan, and M. R. Hasan, "Context-le: Cross-distribution generalization for longitudinal experiential data via narrative-based llm representations," in *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025, pp. 15 335–15 360. (document), 1.1, 1.3, 1.5, 1.6, 1.6, 1.10.4, 1.13, 1.14.5, 2.1, 2.6.4, 2.7.5, 3.1, 3.1, 3.3.4, 3.4.1, 3.4.1, 3.5, 3.3, 3.10, 3.11, 3.10.1, 4, 3.10, 3.10.3, 3.10.3, 2, 3.12, 3.10.5,

3.10.6, 3.13, 3.14, 3.15, 3.16, 3.10.7, 3.17, 3.18, 3.10.8, 4.1, 4.2, 4.6, 6.1, 6.2, 6.4.4, 6.5.1, 6.5.4, 6.6, 6.7.1, 6.7.2, 6.7.4, 6.8.2, 6.8.4, 6.9.2, 6.10.6, 6.11, A.9, A.19, B.1, B.5, B.6, B.10

[4] R. Larson and M. Csikszentmihalyi, "The experience sampling method," *New Directions for Methodology of Social and Behavioral Science*, vol. 15, pp. 41–56, 1983. 1.1, 1.10.1, 6.1

[5] A. A. Stone and S. Shiffman, "Ecological momentary assessment in behavioral medicine," *Annals of Behavioral Medicine*, vol. 16, no. 3, pp. 199–202, 1994. 1.1, 1.10.1, 6.1

[6] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annual Review of Clinical Psychology*, vol. 4, pp. 1–32, 2008. 1.1, 1.10.1, 6.1, 6.9.4

[7] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," *Annual Review of Clinical Psychology*, vol. 13, pp. 23–47, 2017. 1.1, 1.2, 1.10.1, 2.1, 2.4.1, 6.1, 6.8.2, 6.9.1, 6.9.4

[8] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 3–14. 1.1, 1.2, 1.4, 1.10.2, 2.1, 2.4.1, 6.4.1, 6.8.1

[9] R. Wang, P. Hao, X. Zhou, A. T. Campbell, and G. Harari, "Smartgpa: How smartphones can assess and predict academic performance of college stu-

dents," *GetMobile: Mobile Computing and Communications*, vol. 19, no. 4, pp. 13–17, 2016. 1.1, 1.2, 1.4, 1.10.2, 2.1, 2.4.1, 6.4.1, 6.8.1

[10] X. Li, X. Zhu, X. Zhu, Y. Ji, and X. Tang, "Student academic performance prediction using deep multi-source behavior sequential network," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2020, pp. 570–582. 1.1, 1.2, 1.4, 1.10.2, 2.1, 2.4.1, 6.4.1

[11] S. Nepal, W. Liu, A. Pillai, W. Wang, V. Vojdanovski, J. F. Huckins, C. Rogers, M. L. Meyer, and A. T. Campbell, "Capturing the college experience: A four-year mobile sensing study of mental health, resilience and behavior of college students during the pandemic," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 38:1–38:37, March 2024. 1.1, 1.2.4

[12] X. Xu, X. Liu, H. Zhang, W. Wang, S. Nepal, Y. Sefidgar, W. Seo, K. S. Kuehn, J. F. Huckins, M. E. Morris, P. S. Nurius, E. A. Riskin, S. Patel, T. Althoff, A. Campbell, A. K. Dey, and J. Mankoff, "GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 190:1–190:34, Jan. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3569485 1.1, 1.2.4, 5.10.1, 5.11.1

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

USA: Association for Computing Machinery, Sep. 2014, pp. 3–14. [Online]. Available: http://doi.org/10.1145/2632048.2632054 1.1, 1.2.4

[14] 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, "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.*, vol. 28, no. 1, pp. 3:1–3:41, 2021. 1.1, 1.2.4

[15] A. Hayat, B. Khan, and M. R. Hasan, "Leveraging language models for analyzing longitudinal experiential data in education," *arXiv:2503.21617*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2503.21617 1.1, 1.2.4, 5.10.1, 2

[16] X. Xu *et al.*, "Globem: Cross-distribution generalization benchmark for longitudinal behavioral modeling," *arXiv preprint*, 2023, replace with final archival metadata if available. 1.1, 1.2, 1.3, 1.4.4, 1.13, 2.1, 2.4.1, 2.5.5, 2.6.4, 2.2, 2.7.5, 3.3, 3.10.1, 4.1, 4.7.1, 6.1, 6.4.1, 6.8.2, 6.9.1

[17] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1–2, pp. 151–175, 2010. 1.1, 1.3, 1.10.6, 2.1, 2.5.5

[18] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2021. 1.1, 1.3, 1.10.6, 2.5.5

[19] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019. 1.1, 1.8, 1.10.5, 6.4.5, 6.7.5, 6.10.4

[20] A. Hayat, B. Khan, and M. R. Hasan, "Leveraging language models for analyzing longitudinal experiential data in education," *arXiv preprint arXiv:2503.21617*, 2024. 1.1, 1.5, 1.14.2, 2.1, 2.3, 2.6.2, 2.2, 3.1, 3.1, 3.3.2, 3.3.3, 3.8, 3.8.1, 3.8.2, 3.3, 3.3, 3.8.3, 3.9.4, 3.5, 3.9.4, 3.11, 4.1, 4.6, 6.1, 6.4.2, 6.5.1, 6.5.2, 6.6, 6.8.1, 6.11

[21] ——, "Le-viz: Generalizable visuospatial multimodal modeling of longitudinal experiential data," 2026, manuscript / submission. 1.1, 1.3, 1.7, 1.8, 1.8, 1.10.5, 1.13, 1.14.6, 2.5.3, 2.7.2, 3.12.4, 4.1, 4.2, 4.3, 4.5.1, 4.5.2, 4.1, 4.2, 4.5.4, 4.1, 4.6, 4.2, 4.7.1, 4.3, 4.4, 4.5, 4.12.1, 4.14, 5.1, 5.2, 5.8, 3, 6.1, 6.2, 6.4.5, 6.5.1, 6.5.5, 6.6, 6.7.1, 6.7.5, 6.9.2, 6.10.4, 6.10.6, 6.11, A.19, B.1, B.7, B.11

[22] ——, "Frozen language priors as distribution-invariant constraints for behavioral generalization," 2026, neurIPS submission draft. 1.1, 1.3, 1.9, 1.9, 1.10.6, 1.13, 1.14.7, 2.1, 2.5.3, 2.5.5, 2.6.4, 2.2, 2.7.2, 2.7.5, 3.12.4, 5.1, 5.8, 5.10.3, 6.1, 6.2, 6.4.6, 6.5.1, 6.5.6, 6.6, 6.7.1, 6.7.5, 6.7.6, 6.8.2, 6.8.3, 6.8.4, 6.9.2, 6.10.4, 6.10.6, 6.11, A.19, B.1, B.8, B.12

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321–1330. 1.1, 1.3, 1.9, 1.10.6, 1.13, 6.6, 6.8.4, 6.9.4, 6.10.1

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, 2017. 1.1, 1.3, 1.9, 1.10.6, 1.13

[25] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017. 1.1, 1.3, 1.10.6, 1.13

[26] A. Hayat and M. R. Hasan, "Personalization and contextualization of large language models for improving early forecasting of student performance," NeurIPS 2023 Workshop on Generative AI for Education (GAIED), 2023. 1.1, 1.4.3, 1.5, 1.10.2, 1.13, 1.14.1, 2.1, 2.3, 2.4.2, 2.5.1, 2.5.2, 2.6.1, 2.1, 2.2, 3.1, 3.1, 3.2, 3.2, 3.3.3, 3.3.4, 3.4.1, 3.6, 3.6.1, 3.1, 3.6.2, 3.2, 3.3, 3.11, 6.4.2, 6.5.2, 6.7.2, 6.11, A.2, A.2, A.3, A.4, A.5, A.19, B.1, B.2

[27] A. Hayat, B. Khan, and M. Hasan, "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*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 137–148. 1.1, 1.4.3, 1.5, 1.10.2, 1.13, 1.14.1, 2.1, 2.3, 2.6.1, 2.2, 3.1, 3.1, 3.1, 3.2, 3.7, 3.3, 4.6, 6.4.2, 6.5.2, 6.7.2, A.2, A.2, B.1, B.3

[28] G. Yfantidou *et al.*, "Lifesnaps: A longitudinal multimodal dataset for human behavior and well-being," *Scientific Data*, 2022, update fields if you have the exact publication metadata. 1.2, 3.3, 3.10.1, 4.1, 6.8.2, 6.9.1

[29] K. Mundnich, B. M. Booth, M. L'Hommedieu, T. Feng, B. Girault, J. L'Hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte, T. H. Falk, K. Lerman, E. Ferrara, and S. Narayanan, "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, vol. 7, p. 354, 2020. [Online]. Available: https://doi.org/10.1038/s41597-020-00655-3 1.2

[30] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. 1.2.4, 1.6, 1.10.3

[31] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, p. 6085, 2018. 1.2.4, 1.6, 1.10.3, 2.5.4, 6.4.3

[32] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "BRITS: Bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*, 2018. 1.2.4, 1.6, 1.10.3, 2.5.4, 6.4.3

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 1.4.3, 2.5.1

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017. 1.4.3, 1.5, 3.4

[35] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, p. 5?32, Oct. 2001. [Online]. Available: https://doi-org.libproxy.unl.edu/10.1023/A:1010933404324 1.4.3, 2.4.3

[36] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 2020. 1.5

[37] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Davis, D. Demszky, C. Donahue,

M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. Ho, K. Hong, D. Hur, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, P. Liang, M. Majmudar, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, D. Narayanan, B. Newman, A. Nie, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," 2021. 1.5, 1.10.4

[38] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. 1.5

[39] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The Llama 3 herd of models," 2024. 1.5, 3.9.2, 6.8.3, 6.9.3, 6.10.5

[40] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 19 730–19 742. 1.8, 1.10.5

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. 1.9, 3.4.2

[42] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, 2023. 1.9, 3.4.2

[43] J. Herzig, P. Nowak, T. Müller, F. Piccinno, and J. M. Eisenschlos, "TAPAS: Weakly supervised table parsing via pre-training," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4320–4333. 1.10.4

[44] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990. 2.5.1

[45] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734. 2.5.1

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 2.5.2

[47] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018. 2.5.2

[48] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *ArXiv*, vol. abs/2211.14730, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254044221 2.5.3, 1

[49] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," 2024. [Online]. Available: https://arxiv.org/abs/2310.06625 2.5.3, 1

[50] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–20, 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2022.3195549 2.5.5

[51] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2020. [Online]. Available: https://arxiv.org/abs/1907.02893 2.5.5

[52] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "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)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202/ 2.7.1

[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423 2.7.1

[54] A. Hayat, B. Khan, and M. R. Hasan, "Harnessing language models to predict and enhance stem engagement from longitudinal experiential data," in *ASEE Annual Conference and Exposition*, 2025. 3.1, 3.1, 3.4, 3.3, 3.9, 3.9.1, 3.9.2, 3.4, 3.11

[55] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, 2018. 3.3.1

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186. 3.3.1, 3.9.2

[57] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. 3.3.3, 3.4, 3.5.1

[58] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. 3.3.3, 3.5.1, 6.8.3, 6.9.3, 6.10.5

[59] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. 3.3.3, 3.5.1, 6.7.1, 6.8.3, 6.9.3, 6.10.5

[60] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," in *arXiv preprint arXiv:1907.11692*, 2019. 3.9.2

[61] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. 3.9.3, 3.9.3, 6.1, 6.4.3, 6.7.3

[62] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Wiley, 2019. 3.9.3, 3.9.3, 6.1, 6.4.3, 6.7.3

[63] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. 3.9.3

[64] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012. 3.9.3

[65] J. Yoon, J. Jordon, and M. van der Schaar, "Gain: Missing data imputation using generative adversarial nets," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 5689–5698. 3.9.3

[66] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002. 3.9.3

[67] B. Tversky, "Visuospatial reasoning," *The Cambridge Handbook of Thinking and Reasoning*, pp. 209–240, 2005. 4.1, 4.4, 6.4.5

[68] ——, "Visualizing thought," *Topics in Cognitive Science*, vol. 3, no. 3, pp. 499–535, 2011. 4.1, 4.4, 6.4.5

[69] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," 2024, project blog / technical release. 4.5.4

[70] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2023. 4.6, 6.4.1

[71] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2024. 4.6, 6.4.1

[72] N. T. Thach, P. Habecker, A. R. Eisenbraun, W. A. Mason, K. A. Tyler, B. Khan, and H. Chan, "MuHBoost: Multi-label boosting for practical longitudinal human behavior modeling," in *International Conference on Learning Representations*, 2025. 4.6

[73] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," 2024. 4.6

[74] S. Zhong, W. Ruan, M. Jin, H. Li, Q. Wen, and Y. Liang, "Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting," 2025. 4.6

[75] A. Hayat, B. Khan, and M. R. Hasan, "ConText-LE: Cross-distribution generalization for longitudinal experiential data via narrative-based LLM representations," in *Findings of the Association for Computational*

Linguistics: EMNLP 2025, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 15 335–15 360. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.830/ 5.1, 5.2, 5.4, 5.6.3, 5.9.5, 2, 5.11.1

[76] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," 2017. [Online]. Available: https://arxiv.org/abs/1705.09406 5.1, 5.2

[77] M. Y. Sim, W. E. Zhang, X. Dai, and B. Fang, "Can VLMs actually see and read? a survey on modality collapse in vision-language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 24 452–24 470. [Online]. Available: https://aclanthology.org/2025.findings-acl.1256/ 5.1, 5.2

[78] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, " Balanced Multimodal Learning via On-the-fly Gradient Modulation ," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 8228–8237. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00806 5.1

[79] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000. [Online]. Available: https://arxiv.org/abs/physics/0004057 5.1

[80] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5. 5.1

[81] M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021. 5.2

[82] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annual Review of Clinical Psychology*, vol. 4, pp. 1–32, 2008. 5.3

[83] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," *Annual Review of Clinical Psychology*, vol. 13, pp. 23–47, 2017. 5.3

[84] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," 2018. [Online]. Available: https://arxiv.org/abs/1705.07115 5.4, 5.9.4

[85] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1999. 5.6.2, 5.15

[86] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61 – 78, 1998. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/79/1/1520-0477_1998_079_0061_apgtwa_2_0_co_2.xml 5.6.2

[87] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020 5.6.2

[88] S. Yfantidou, C. Karagianni, S. Efstathiou *et al.*, "Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild," *Scientific Data*, vol. 9, no. 1, p. 663, 2022. [Online]. Available: https://doi.org/10.1038/s41597-022-01764-x 5.10.1

[89] A. Hayat, B. Khan, and M. Hasan, "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)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 137–148. [Online]. Available: https://aclanthology.org/2024.bea-1.13/ 5.10.1, 2

[90] N. T. Thach, P. Habecker, A. R. Eisenbraun, W. A. Mason, K. A. Tyler, B. Khan, and H. Chan, "Muhboost: Multi-label boosting for practical longitudinal human behavior modeling," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025, accepted. Available at https://openreview.net/pdf?id=BAelAyADqn. 2

[91] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," 2024. [Online]. Available: https://arxiv.org/abs/2401.06866 2

[92] S. Zhong, W. Ruan, M. Jin, H. Li, Q. Wen, and Y. Liang, "Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting," 2025. [Online]. Available: https://arxiv.org/abs/2502.04395 3

[93] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/ 3

Appendix A

Additional Results

A.1 Purpose of This Appendix

The main chapters of this dissertation focused on the central methodological arc and the most consequential empirical findings. This appendix expands that record. It preserves additional comparisons, ablations, reverse-direction transfer results, model-family analyses, and implementation-oriented details that support the core argument but would have interrupted the flow of the core narrative. The goal here is not to introduce new claims. Rather, it is to document more fully the empirical basis for the claims already advanced in Chapters 2 through 5.

Across the dissertation, six linked ideas recur: traditional numerical models are limited for LE forecasting; contextual language modeling already provides a substantial improvement; missingness should be treated as meaningful structure; narrative formulation improves cross-distribution transfer; multimodal visual-text modeling preserves structure lost by pure serialization; and frozen multimodal constraints produce the strongest and most reliable generalization. The additional results collected here reinforce each of those claims with broader quantitative evidence and with more detailed views of the design choices behind them.

A.2 Supplementary Results for Early Forecasting and Classical Baselines

The earliest LLM stage of the dissertation asked whether an instruction-tuned language model could forecast future academic outcomes from short, partially observed student trajectories better than conventional numerical architectures. The main chapters summarized that transition conceptually. Tables A.1 and A.2 preserve two of the most informative empirical views from that phase [26, 27].

Table A.1: Representative early forecasting progression across observation windows. "Best numeric baseline" corresponds to the strongest conventional baseline reported for the same split, while the FLAN-T5 rows use the richest contextualized textual formulation. Accuracies are reported in percent.

Method	8-week	4-week	2-week
Best numeric baseline (SVM)	68	59	59
FLAN-T5 Small	82	75	64
FLAN-T5 Base	86	84	68
FLAN-T5 Large	**89**	**84**	**77**

Table A.2: Context and personalization ablation using FLAN-T5 Large. C = cognitive, NC = non-cognitive, and BG = background/distal factors. Accuracies are reported in percent [26, 27].

Representation	8-week	4-week	2-week
C + NC + BG	**89**	**84**	**77**
C + NC	82	77	68
C + BG	77	77	64
C only	73	70	52

These tables are useful because they expose two early facts that remain important for the rest of the dissertation. First, the advantage of contextual language modeling appears most clearly in the short-window regime, where the available evidence is most incomplete. Second, the gain does not come only from

converting numbers into words. It comes from preserving cognitive, non-cognitive, and background structure together. Context functions as predictive evidence rather than as decorative metadata.

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

[26]

Features	Class	8-week				4-week				2-week			
		P	R	F1	A	P	R	F1	A	P	R	F1	A
C + NC + D	AR	1.00	0.86	0.92	**0.86**	0.78	1.00	0.88	**0.84**	0.55	0.86	0.67	**0.68**
	PR	0.78	0.70	0.74		0.89	0.80	0.84		0.71	0.50	0.59	
	AV	0.91	0.91	0.91		0.79	1.00	0.88		0.71	0.91	0.80	
	OU	0.83	0.94	0.88		0.92	0.69	0.79		0.75	0.56	0.64	
C + NC	0.88	1.00	1.00	1.00	0.80	0.71	0.71	0.71	0.73	0.46	0.86	0.60	0.61
	0.58	0.70	0.70	0.70		0.75	1.00	0.67		0.64	0.70	0.67	
	0.82	0.71	0.91	0.80		0.69	1.00	0.81		0.67	0.73	0.70	
	0.77	0.85	0.69	0.76		0.77	0.62	0.69		0.75	0.38	0.50	
C + D	0.60	0.78	1.00	0.88	0.75	1.00	1.00	1.00	0.73	0.67	0.86	0.75	0.64
	0.83	0.78	0.70	0.74		0.69	0.90	0.78		0.56	0.50	0.53	
	0.73	0.73	0.73	0.73		0.64	0.82	0.72		0.86	0.55	0.67	
	0.71	0.73	0.69	0.71		0.70	0.44	0.54		0.58	0.69	0.63	
C	0.64	0.64	1.00	0.78	0.70	0.86	0.86	0.86	0.66	0.50	0.57	0.53	0.48
	0.67	0.75	0.60	0.67		0.57	0.40	0.47		0.83	0.50	0.62	
	0.73	0.82	0.82	0.82		0.53	0.82	0.64		0.35	0.64	0.45	
	0.54	0.64	0.56	0.60		0.77	0.62	0.69		0.50	0.31	0.38	

A.3 Supplementary Results for Missingness-Aware Representation

The main missingness chapter focused on the conceptual argument for treating absence as meaningful structure and on the central CRILM results. The source appendix contains a different kind of evidence: supporting tables that clarify dataset coverage, descriptor construction, and the classical imputation settings

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

[26]

Features	Class	8-week				4-week				2-week			
		P	R	F1	A	P	R	F1	A	P	R	F1	A
C + NC + D	AR	1.00	0.86	0.92	**0.82**	0.60	0.43	0.50	**0.75**	0.60	0.86	0.71	**0.64**
	PR	1.00	0.40	0.57		0.89	0.80	0.84		0.62	0.50	0.56	
	AV	0.77	0.91	0.83		0.77	0.91	0.83		0.67	0.73	0.70	
	OU	0.76	1.00	0.86		0.71	0.75	0.73		0.64	0.56	0.60	
C + NC	0.88	1.00	0.93	0.82	0.75	0.50	0.71	0.59	0.66	0.42	0.71	0.53	0.59
	0.58	0.70	0.64	0.75		0.67	0.60	0.63		0.80	0.40	0.53	
	0.82	0.82	0.82	0.85		0.67	0.73	0.70		0.50	0.73	0.59	
	0.77	0.62	0.69	0.83		0.77	0.62	0.69		0.82	0.56	0.67	
C + D	0.60	0.86	0.71	0.88	0.70	0.86	0.86	0.86	0.64	0.67	0.86	0.75	0.59
	0.83	0.50	0.62	0.84		0.75	0.90	0.82		0.60	0.60	0.60	
	0.73	0.73	0.73	0.70		0.44	0.64	0.52		0.56	0.82	0.67	
	0.71	0.75	0.73	0.73		0.67	0.38	0.48		0.56	0.31	0.40	
C	0.64	1.00	0.78	0.71	0.64	0.67	0.57	0.62	0.61	0.25	0.43	0.32	0.41
	0.67	0.60	0.63	0.71		0.50	0.30	0.37		0.40	0.20	0.27	
	0.73	0.73	0.73	0.69		0.64	0.82	0.72		0.45	0.45	0.45	
	0.54	0.44	0.48	0.79		0.61	0.69	0.65		0.50	0.50	0.50	

Table A.5: Evaluation of the three baseline models trained with cognitive features using the 8-week, 4-week, and 2-week datasets. The best results are in **bold**.
Legends: AR=At-Risk, PR=Prone-To-Risk, AV=Average, OU=Outstanding, P=Precision, R=Recall, F1=F1 Score, A=Accuracy

[26]

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

against which CRILM was compared. These materials are retained here because they support reproducibility without duplicating the main chapter's headline results [1].

Table A.6: Dataset summary for the UCI benchmarks used in the CRILM appendix analysis. N denotes the number of instances and d denotes the number of features [1].

Dataset	N	d	Description
Iris	150	4	The dataset contains 3 classes of 50 instances each, referring to types of iris plants.
Wine	178	13	Results of a chemical analysis of wines grown in Italy, with three types represented.
Seeds	210	7	Properties of three varieties of wheat: Kama, Rosa, and Canadian.
Glass Identification	214	9	Classification of types of glass for criminological investigation.
Ionosphere	351	34	Phased array of 16 high-frequency antennas, targeting free electrons in the ionosphere.
Breast Cancer Wisconsin	569	30	Binary classification from digitized images of a fine needle aspirate of breast masses.

Table A.8: Optimal k values for k-NN imputation across MCAR, MAR, and MNAR missingness patterns using Llama and FLAN-T5 models on six datasets [1].

Dataset	**Missing pattern**	**Model**	k	**Acc. (%)**
Iris	MCAR	Llama	5	84.60
Iris	MCAR	FLAN-T5	5	78.80
Iris	MAR	Llama	5	87.40
Iris	MAR	FLAN-T5	3	82.20
Iris	MNAR	Llama	7	76.60
Iris	MNAR	FLAN-T5	5	72.20
Wine	MCAR	Llama	3	80.20
Wine	MCAR	FLAN-T5	3	74.60
Wine	MAR	Llama	5	86.20
Wine	MAR	FLAN-T5	5	82.40
Wine	MNAR	Llama	5	73.20

Table A.8 (continued)

Dataset	Missing pattern	Model	k	Acc. (%)
Wine	MNAR	FLAN-T5	3	71.60
Seeds	MCAR	Llama	3	79.40
Seeds	MCAR	FLAN-T5	3	79.00
Seeds	MAR	Llama	3	81.60
Seeds	MAR	FLAN-T5	5	81.20
Seeds	MNAR	Llama	3	72.20
Seeds	MNAR	FLAN-T5	5	71.60
Glass	MCAR	Llama	5	52.40
Glass	MCAR	FLAN-T5	5	44.20
Glass	MAR	Llama	5	57.60
Glass	MAR	FLAN-T5	5	49.20
Glass	MNAR	Llama	3	41.40
Glass	MNAR	FLAN-T5	5	36.60
Ionosphere	MCAR	Llama	5	86.80
Ionosphere	MCAR	FLAN-T5	5	87.20
Ionosphere	MAR	Llama	5	85.80
Ionosphere	MAR	FLAN-T5	5	83.40
Ionosphere	MNAR	Llama	3	79.60
Ionosphere	MNAR	FLAN-T5	5	78.20
Cancer	MCAR	Llama	5	85.20
Cancer	MCAR	FLAN-T5	3	83.00
Cancer	MAR	Llama	5	89.80
Cancer	MAR	FLAN-T5	5	85.60
Cancer	MNAR	Llama	5	82.40
Cancer	MNAR	FLAN-T5	5	78.40

Table A.7: Corrected feature-specific contextually relevant descriptors for three selected datasets. This table replaces the earlier erroneous version and follows the source appendix directly [1].

Dataset	Features containing Missing values	Descriptors of missing values
Iris	1. Sepal Length	1. Sepal Length: Unavailable
	2. Sepal Width	2. Sepal Width: Unavailable
	3. Petal Length	3. Petal Length: Unavailable
	4. Petal Width	4. Petal Width: Unavailable
Wine	1. Alcohol	1. Alcohol content not provided for this wine sample.
	2. Malic acid	2. Malic acid quantity missing for this wine sample.
	3.Ash	3. Ash content data not recorded for this wine sample.
	4. Alcalinity of ash	4. Alcalinity of ash information unavailable for this wine sample.
	5. Magnesium	5. Magnesium level not specified for this wine sample.
	6. Total phenols	6. Total phenols data missing for this wine sample.
	7. Flavanoids	7. Flavanoids content not available for this wine sample.
	8. Nonflavanoi phenols	8. Nonflavanoid phenols quantity not provided for this wine sample.
	9. Proanthocyanins	9. Proanthocyanins data missing for this wine sample.
	10. Color Intensity	10. Color intensity information not recorded for this wine sample.
	11. Hue	11. Hue value not specified for this wine sample.
	12.OD280/OD315 of diluted wines	12. OD280/OD315 data missing for this wine sample.
	13. Proline	13. Proline content not available for this wine sample
Seeds	1. Area	1. Kernel area not provided.
	2. Perimeter	2. Kernel perimeter information missing.
	3. Compactness	3. Kernel compactness data not recorded.
	4. Length of kernel	4. Length of kernel data missing.
	5. Width of kernel	5. Width of kernel data missing.
	6. Asymmetry coefficient	6. Asymmetry coefficient information missing.
	7. Length of kernel groove	7. Length of kernel groove information missing.

A.4 Supplementary Statistical Validation for ConText-LE

The ConText-LE chapter already presented the full forward and reverse transfer tables in the main dissertation text. Those are therefore omitted here to avoid redundancy. What remains especially useful from the source appendix is the direct statistical validation of the Meta-Narrative advantage. Table A.9 preserves the paired significance analysis comparing Meta-Narrative to the other three textual representations on a per-instance correctness basis.

This statistical view complements the raw accuracy and F1 results reported in Chapter 3. It shows that the gains from Meta-Narrative are not merely anecdotal improvements on a few settings, but persist strongly enough to appear in paired comparisons across datasets and test regimes.

Table A.9: Pairwise t-test results comparing Meta-Narrative with Complete Sequence, Statistical Summary, and Natural Language String. Asterisks denote uncorrected significance levels: * indicates $p < 0.05$ and ** indicates $p < 0.001$ [3].

Dataset	Model Compared To	ID	OOD
GLOBEM	Complete Sequence	**	**
	Statistical Summary	*	**
	Natural Language String	*	**
LifeSnaps	Complete Sequence	*	**
	Statistical Summary	–	*
	Natural Language String	–	**
MFAFY	Complete Sequence	*	*
	Statistical Summary	*	*
	Natural Language String	*	*

A.5 Supplementary Results for PRISM

The final framework, PRISM, integrates temporal, visual, and semantic evidence while constraining adaptation through frozen priors and a dual-path learning objective. The main chapter focused on the architecture and its central results. The source appendix, however, includes more granular path-specific ablations than were practical to include in the main text. These detailed tables are retained here because they show how the relative behavior of PRISM changes depending on the inference path and on which modality is removed.

Table A.10: Progression of the strongest OOD accuracy across major dissertation stages. The values highlight the cumulative effect of representational and training-constraint advances.

Dataset	Traditional	Text-only LLM	LE-Viz	PRISM$_{gen}$
GLOBEM	51.06	67.40	72.86	**79.93**
LifeSnaps	48.44	67.19	71.88	**81.25**
MFAFY	50.57	64.86	66.57	**73.71**

The next four tables unpack the dual-path objective more completely than the consolidated summary in Chapter 5. Rather than only reporting the best setting, they show how predictive-only, generative-only, fixed-weight, and learned-weight training behave separately on both the generation and prediction paths.

Table A.11: Dual-path ablation: OOD results, Generation path (Acc / P / R / F1, %).

	GLOBEM				LifeSnaps				MFAFY			
Objective	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
$\mathcal{L}_{\text{pred}}$ only	59.12	62.79	57.20	53.00	64.06	66.55	63.39	61.97	55.71	67.81	63.65	54.67
$\mathcal{L}_{\text{gen}}$ only	71.33	77.39	72.64	70.36	70.31	71.97	69.84	69.41	68.29	67.46	61.24	60.79
Fixed $\alpha = 0.5$	75.73	80.17	74.53	74.15	76.56	80.23	76.00	75.55	71.71	74.63	64.29	64.12
Learned	**79.93**	**82.53**	**79.06**	**79.14**	**81.25**	**83.46**	**80.84**	**80.78**	**73.71**	**78.03**	**66.50**	**66.77**

Table A.12: Dual-path ablation: ID results, Generation path (Acc / P / R / F1, %).

	GLOBEM				LifeSnaps				MFAFY			
Objective	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
$\mathcal{L}_{\text{pred}}$ only	61.43	66.05	60.52	57.35	68.75	67.95	61.67	61.35	58.49	69.90	65.68	57.75
$\mathcal{L}_{\text{gen}}$ only	74.44	79.30	73.78	72.97	75.00	75.00	70.00	70.91	73.58	76.05	66.97	67.46
Fixed $\alpha = 0.5$	77.58	79.21	77.19	77.08	81.25	83.33	85.00	81.18	73.58	72.63	73.86	72.80
Learned	**81.17**	**83.39**	**80.75**	**80.69**	**87.50**	**87.50**	**90.00**	**87.30**	**75.47**	**74.18**	**72.42**	**73.01**

Table A.13: Dual-path ablation: OOD results, Prediction path (Acc / P / R / F1, %).

	GLOBEM				LifeSnaps				MFAFY			
Objective	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
$\mathcal{L}_{\text{pred}}$ only	65.10	73.70	66.71	62.85	64.06	66.55	63.39	61.97	64.57	61.50	56.77	55.21
$\mathcal{L}_{\text{gen}}$ only	55.86	67.08	58.03	50.27	64.06	71.47	63.10	59.73	52.29	66.55	60.65	50.47
Fixed $\alpha = 0.5$	**68.26**	**75.22**	**66.60**	**64.57**	**70.31**	**76.04**	**69.55**	**68.06**	**67.43**	**65.71**	**60.70**	**60.34**
Learned	67.08	73.49	65.40	63.22	68.75	74.97	67.94	66.07	65.71	62.82	59.63	59.40

Table A.14: Dual-path ablation: ID results, Prediction path (Acc / P / R / F1, %).

	GLOBEM				LifeSnaps				MFAFY			
Objective	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
$\mathcal{L}_{\text{pred}}$ only	**72.20**	**75.09**	**71.63**	**71.02**	**75.00**	75.00	**76.67**	**74.60**	**71.70**	**72.08**	65.45	65.81
$\mathcal{L}_{\text{gen}}$ only	56.95	63.23	57.95	52.68	62.50	58.33	56.67	56.36	56.60	69.00	64.17	55.59
Fixed $\alpha = 0.5$	72.65	75.43	72.10	71.54	**81.25**	**80.91**	78.33	**79.22**	71.70	70.18	67.42	68.01
Learned	70.85	76.39	70.10	68.74	75.00	**85.71**	66.67	66.67	69.81	67.74	**65.91**	**66.35**

Note on prediction-path trade-offs. The source appendix observes that fixed equal weighting can slightly outperform learned weighting on the prediction path under OOD evaluation, even though learned weighting is best for the generation path. This is consistent with the final framework's design choice: PRISM optimizes

for the stronger generative inference route, while still preserving a viable predictive branch as a complementary decision path.

The modality-removal tables below provide the full path-specific ablation record. These are more detailed than the condensed modality-removal summary in Chapter 5 and are therefore useful for understanding how the importance of text, vision, and temporal structure shifts between the generation and prediction routes.

Table A.15: Modality removal: OOD results, Generation path (Acc / P / R / F1, %).

	GLOBEM				**LifeSnaps**				**MFAFY**			
Config.	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Full	**79.93**	**82.53**	**79.06**	**79.14**	**81.25**	**83.46**	**80.84**	**80.78**	**73.71**	**78.03**	**66.50**	**66.77**
w/o Text	65.62	72.93	66.47	63.33	68.75	74.97	69.50	67.18	61.14	70.96	67.76	60.73
w/o Vision	68.81	75.61	67.17	65.32	71.88	74.23	72.34	70.85	64.29	72.45	70.28	64.10
w/o Temporal	71.13	77.20	69.64	68.40	73.44	77.95	74.05	72.63	67.71	65.44	62.28	62.45

Table A.16: Modality removal: ID results, Generation path (Acc / P / R / F1, %).

	GLOBEM				**LifeSnaps**				**MFAFY**			
Config.	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Full	**81.17**	**83.39**	**80.75**	**80.69**	**87.50**	**87.50**	**90.00**	**87.30**	**75.47**	74.18	**72.42**	**73.01**
w/o Text	67.71	71.60	67.00	65.62	75.00	73.33	73.33	73.33	64.15	72.55	70.23	63.95
w/o Vision	69.96	76.40	69.15	67.44	81.25	80.91	78.33	79.22	67.92	65.53	63.41	63.74
w/o Temporal	72.65	78.15	71.93	70.80	81.25	80.16	81.67	80.57	71.70	70.87	66.44	67.00

Table A.17: Modality removal: OOD results, Prediction path (Acc / P / R / F1, %).

	GLOBEM				**LifeSnaps**				**MFAFY**			
Config.	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Full	**67.08**	**73.49**	**65.40**	**63.22**	**68.75**	**74.97**	**67.94**	**66.07**	**65.71**	62.82	59.63	59.40
w/o Text	61.10	69.17	58.98	53.93	60.94	65.62	61.73	58.67	58.86	69.88	65.93	58.20
w/o Vision	61.94	71.13	63.73	59.01	59.38	67.07	58.26	52.73	60.86	70.83	67.53	60.41
w/o Temporal	62.98	72.42	64.75	60.20	62.50	68.57	63.34	60.00	61.43	71.10	67.99	61.04

Table A.18: Modality removal: ID results, Prediction path (Acc / P / R / F1, %).

	GLOBEM				LifeSnaps				MFAFY			
Config.	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Full	**70.85**	**76.39**	**70.10**	**68.74**	75.00	**85.71**	66.67	66.67	**69.81**	67.74	65.91	**66.35**
w/o Text	63.23	71.99	64.18	60.04	68.75	66.36	65.00	65.37	60.38	70.79	67.20	59.86
w/o Vision	63.68	71.93	62.67	58.90	68.75	67.95	61.67	61.35	62.26	62.04	62.80	61.59
w/o Temporal	65.02	72.97	64.06	60.82	68.75	77.27	75.00	68.63	64.15	70.16	69.24	64.10

Modality contribution on the prediction path. The contribution ordering on the prediction path is broadly consistent with the generation path: removing text typically causes the largest OOD drop on GLOBEM and LifeSnaps, while on MFAFY the three removals produce more comparable degradations. This consistency across inference routes provides additional confidence that the modality ordering reflects information content rather than an artifact of the generative decoding mechanism.

A.6 Variance, Stability, and Reproducibility Notes

Because the dissertation ultimately concerns reliable modeling rather than one-off peak scores, a concise implementation-oriented summary is useful. The table below preserves representative settings from the major stages of the dissertation. These details matter because LE datasets are comparatively small and can be sensitive to adapter rank, decoding policy, and whether the backbone is frozen.

A.7 Concluding Remarks

Taken together, the supplementary results in this appendix reinforce the main empirical argument of the dissertation. The move from traditional baselines to contextual language modeling matters. The treatment of missingness matters.

Table A.19: Representative implementation details across major dissertation stages. The purpose is to preserve the settings most relevant for reproducibility and interpretation.

Stage	**Base model**	**Adaptation / decoding**	**Representative details**
Early LLM forecasting	FLAN-T5 Small / Base / Large	full fine-tuning	batch size 6, 50 epochs, AdamW, 512-token context limit, and observation windows of 2, 4, and 8 weeks [26].
ConText-LE	Llama 3.1 8B Instruct	LoRA + narrative decoding	rank 16, alpha 32, dropout 0.05; temperature 0.7; top-p 1.0; max generation length 300; frequency penalty 0.5 [3].
LE-Viz text-only baselines	Llama 3.1 8B Instruct	LoRA fine-tuning	rank 32, alpha 16, dropout 0.1; learning rate 10^{-5}; batch size 8; 20 epochs [21].
LE-Viz multimodal	LLaVA-NeXT 7B	LoRA fine-tuning	rank 64, alpha 128, dropout 0.05; learning rate 10^{-4}; batch size 4; 20 epochs; greedy decoding with max generation length 2000 [21].
PRISM	frozen multimodal backbones + trainable fusion	dual-path objective	learned homoscedastic weighting over predictive and generative losses; frozen semantic prior used as a constraint rather than fully co-adapted supervision [22].

Narrative abstraction matters. Multimodal structure preservation matters. And disciplined, frozen-prior multimodal learning matters most of all. The additional tables and figures do not alter the overall story; they make it more complete and more auditable.

Appendix B

Prompts, Templates, and Representation Procedures

B.1 Purpose of This Appendix

A central theme of this dissertation is that representation design is part of the method rather than a peripheral implementation detail. In the language-based and multimodal stages of the dissertation, LE data is repeatedly transformed into textual descriptions, narrative abstractions, and visual-text inputs before forecasting occurs. Those transformations materially affect what the model can learn, what kinds of reasoning it can perform, and how well it generalizes under shift. This appendix therefore documents the prompt families and representation procedures that support the major stages of the dissertation [1–3, 21, 22, 26, 27].

B.2 Template Family I: Early Structured Textualization

The earliest language-model stage of the dissertation converted structured student records into natural-language sequences so that forecasting could be posed as instruction-following generation rather than purely numerical classification [26]. The original template is reproduced below with light formatting edits.

[title=Template A: Early structured trajectory description, breakable] **Input**

Table B.1: Prompt families across dissertation stages. Prompt design evolves because the representational goal evolves.

Stage	Representation	Prompt purpose	Primary use
Early forecasting	structured-to-text verbalization	convert short educational trajectories into instruction-ready language	end-of-semester academic forecasting
Contextualized forecasting	richer educational prompt	integrate distal, cognitive, and non-cognitive context	early academic outcome prediction
Missingness-aware modeling	descriptor generation	preserve the semantics of absence rather than hide it	CRILM and qualitative LE forecasting
ConText-LE	Meta-Narrative input + prospective narrative output	abstract trajectories into behaviorally meaningful narrative form	cross-distribution generalization
LE-Viz	text + visual interleaving	coordinate semantic abstraction with spatial structure preservation	multimodal LE forecasting
PRISM	frozen coherence target	convert narrative supervision into a stable semantic constraint	constrained multimodal transfer

Sequence:

A student obtained the following assessment scores in an introductory programming course on [NAME OF LANGUAGE] in [SEMESTER] from week 1 to week [n] for [LIST OF GRADED COMPONENTS]: in week 1, scored [?] out of [?] in [NAME OF COGNITIVE TEST], ..., [MEASURE OF TWO EMOTIONAL ENGAGEMENT FEATURES:] student believes that student might get [X] grade and student is [Y] satisfied with performance; in week 2 [CONTINUE AS BEFORE] ... in week [n], scored [?] out of [?] in [NAME OF COGNITIVE TEST], ..., [MEASURE OF TWO EMOTIONAL ENGAGEMENT FEATURES:] student believes that student might get [X] grade and student is [Y] satisfied with performance. Some background information about the student: Student is a [RACE], [GENDER] in his/her class standing year with a major in [Z].

His/Her family income is [$].

Output Sequence:

leftmargin=2em If the student's grade is (A+, A, A-), output: *At the end of the semester, the student will exhibit an outstanding performance.*

leftmargin=2em If the student's grade is (B+, B, B-), output: *At the end of the semester, the student will exhibit an average performance.*

leftmargin=2em If the student's grade is (C+, C, C-), output: *At the end of the semester, the student will be prone to risk.*

leftmargin=2em If the student's grade is below C-, output: *At the end of the semester, the student will be at-risk.*

This template is representationally simple compared with later chapters, but it establishes three ideas that remain central to the dissertation: temporal order should be explicit, background matters, and the target can be framed in human-readable language rather than as a raw class index.

B.3 Template Family II: Contextualized Educational Forecasting

The next stage enriched the verbalization by integrating distal, proximal cognitive, and proximal non-cognitive information more deliberately [27]. A representative contextualized prompt family can be summarized as follows.

[title=Template B: Contextualized educational forecasting prompt, breakable] You are given a student's background, recent academic performance, and recent non-cognitive responses. Use this information to forecast the student's future academic outcome.

Background: [distal factors]

Recent cognitive trajectory: [weeks 1 to w]

Recent non-cognitive trajectory: [weeks 1 to w]

Forecast the student's likely end-of-semester performance.

The conceptual importance of this family is that context is no longer appended as an afterthought. It becomes part of the main interpretive frame, which is precisely why the contextualized models outperform cognitive-only verbalizations in the short-window forecasting results.

B.4 Template Family III: Missingness-Aware Descriptor Generation

CRILM and the later qualitative engagement work treat missingness as representable structure [1, 2]. In CRILM, one prompt family is used to generate descriptors for missing attributes, and another uses those descriptors during downstream prediction.

[title=Template C: CRILM descriptor-generation instruction, breakable] For any missing attribute values, suggest contextually relevant descriptors to fill in the missing data.

[title=Template D: CRILM downstream prediction instruction, breakable] Predict the class based on the given measurements. Use the context provided by the missing value descriptors to inform the prediction.

A domain-adapted qualitative educational version follows the same logic.

[title=Template E: Educational missingness-aware prompt, breakable] You are given part of a student's longitudinal qualitative trajectory. One response

is missing. Use the surrounding observed information to describe that missing response in a way that preserves context without inventing unsupported detail.

Observed context: `[nearby values, related features, recent trajectory]`
Missing feature and time step: `[feature name / week]`

Generate a concise natural-language descriptor for the missing response.

The feature-selection stage in the three-tier framework also uses prompting, but now the goal is not prediction; it is to identify which qualitative features are most relevant for the target construct.

[title=Template F: Feature-selection prompt, breakable] Given features [Q1: motivation, . . . , Q10: identity] and target [Lecture Engagement shift], select the most predictive subset.

B.5 Template Family IV: ConText-LE Input Representations

ConText-LE compares four textual input representations and studies which one transfers best under distribution shift [3]. The exact system prompts used in that work are important enough to preserve directly.

[title=System Prompt - Statistical Summary, breakable] 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.

[title=System Prompt - Complete Sequence, breakable] 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.

[title=System Prompt - Natural Language String, breakable] 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.

[title=System Prompt - Meta-Narrative, breakable] 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:

leftmargin=2em Identify and describe specific behavioral trends, such as shifts in confidence, motivation, or engagement.

leftmargin=2em Reference specific weeks (e.g., “In Week 1...”, “By Week 3...”).

leftmargin=2em 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”.

leftmargin=2em Avoid vague terms like “overall” or “in general” to ensure analytical precision.

leftmargin=2em Provide a concise, natural, and evidence-based analysis in 3-4 sentences.

leftmargin=2em Exclude any personal or identifying information from the response.

The Meta-Narrative prompt is the most important member of this family because it explicitly asks for behavioral interpretation rather than direct restatement. This is the point in the dissertation where the model is first invited to operate on a semantically grounded abstraction of the trajectory.

B.6 Template Family V: Prospective Narrative Generation and Label Extraction

ConText-LE further shows that output formulation matters as much as input representation. The strongest text-only setting asks the model to generate a future-oriented expert narrative and then extract the binary label from that generated reasoning [3].

[title=System Prompt - Prospective Narrative Generation, breakable] 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.

[title=System Prompt - Prediction Extraction, breakable] 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`.

This pair of prompts matters because it separates narrative interpretation from label extraction. That decomposition later becomes even more meaningful in PRISM, where narrative supervision is converted into a stable semantic constraint rather than used only as a free-form intermediate.

B.7 Template Family VI: Multimodal Prompting in LE-Viz

In LE-Viz, the prompt no longer carries the entire representational burden. The text channel contributes semantic abstraction, while the visual channel preserves temporal and structural geometry [21]. The resulting prompt family can therefore remain semantically focused.

[title=Template G: Multimodal visual-text forecasting prompt, breakable] You are given a visual representation of a longitudinal trajectory and a textual narrative describing that trajectory. Use the visual structure to interpret temporal and feature-wise patterns, and use the text to interpret the broader behavioral meaning. Then generate a concise future-state narrative.

Text summary: `[Meta-Narrative]`

Visual input: `[chart / heatmap / interleaved visual chain]`

The important change is procedural rather than purely lexical. In earlier stages, text had to compensate for structure lost during serialization. In LE-Viz, text can remain semantically focused because the visual representation retains the geometry of the trajectory.

B.8 Template Family VII: Frozen Semantic Targets in PRISM

PRISM does not introduce a wholly unrelated prompt family. Its more important contribution is to change the role of the narrative target [22]. A generated future-state narrative is no longer just an output formulation. Under a frozen prior, it becomes a stable semantic target against which predictive learning is constrained.

[title=Template H: Coherence-target narrative, breakable] You are given a LE trajectory. Generate a semantically grounded description of the trajectory's likely future state that captures its broader behavioral meaning. This narrative should reflect stable, interpretable behavioral structure rather than a narrow task-specific shortcut.

Trajectory: `[structured trajectory or narrative summary]`

This distinction is conceptually important. In ConText-LE, narrative is the output formulation. In PRISM, narrative becomes part of the constraint mechanism that shapes representation learning.

B.9 Dataset-Specific Example Styles

The prompt families above are organized by methodological stage, but the dissertation also used phrasing styles adapted to the main dataset families.

B.9.1 Educational LE style (MFAFY-like)

[title=Example style: educational LE trajectory, breakable] Background: first-year student in an introductory STEM course.

Recent cognitive trajectory:

Week 1: `[performance signals]`

Week 2: `[performance signals]`

Week 3: `[performance signals]`

Week 4: `[performance signals]`

Recent non-cognitive trajectory:

Week 1: `[motivation, confidence, engagement]`

Week 2: `[motivation, confidence, engagement]`

Week 3: `[motivation, confidence, engagement]`

Week 4: `[motivation, confidence, engagement]`

Write a narrative describing the student's recent trajectory and likely next state.

B.9.2 Behavioral sensing style (GLOBEM-like)

[title=Example style: behavioral sensing trajectory, breakable] Recent behavioral trajectory includes mobility pattern, phone usage pattern, sleep-related pattern, social interaction signal, and activity variability. Summarize the trajectory and generate a future-oriented narrative describing the most likely next state.

B.9.3 Wearable well-being style (LifeSnaps-like)

[title=Example style: wearable well-being trajectory, breakable] Recent trajectory includes wearable-derived activity signals, physiological indicators, self-reported well-being cues, and short-term temporal fluctuations. Write a concise narrative describing the recent state and likely next behavioral or well-being outcome.

B.10 Template Family VI-A: Illustrative Examples of ConText-LE Textualizations

The ConText-LE source appendix includes simplified examples showing how the same raw LE window can be verbalized in four different ways. These are useful because they make the representational differences concrete rather than purely conceptual [3].

[title=Example A: Complete Sequence style, breakable] *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.*

[title=Example B: Statistical Summary style, breakable] *Statistical summary over the k-week period: Steps: {"avg": 1050, "std": 350, "min": 500, "max": 1500}. 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}.*

[title=Example C: Natural Language String style, breakable] *Steps: ["500", "1200", "300", "800", ..., "1500", ...]. Sleep Duration: ["7.0", "8.5", "4.0", "7.8", ...,*

"7.2", ...]. Mood: ["3", "3", "5", "4", ..., "3", ...].

[title=Example D: Meta-Narrative style, breakable] *Over the past k weeks, the user's activity levels showed moderate fluctuation with an overall increasing trend toward 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 show why the four representations behave differently. Complete Sequence preserves detail but is verbose, Statistical Summary compresses aggressively, Natural Language String preserves ordering with limited interpretation, and Meta-Narrative explicitly encodes behavioral meaning.

B.11 Template Family VI-B: LE-Viz Construction and Multimodal Assembly Details

The LE-Viz appendix clarifies that prompt wording alone does not define the multimodal method. Construction choices for the visual channel are also part of the representation procedure [21].

The practical lesson is that multimodal LE modeling depends not only on what is said in the prompt, but also on how the trajectory is spatially arranged before the VLM ever receives it.

B.12 Template Family VII-A: PRISM as Prompt-Constrained Supervision

PRISM changes the role of narrative supervision from free-form explanation to a frozen coherence target. The source appendix does not introduce a wholly

Table B.2: LE-Viz multimodal construction details retained from the source appendix in condensed form.

Component	Representative design choices
Feature trajectory line charts	640×480 pixels; x-axis corresponds to days in the k-week window; y-axis reflects the feature range; missing values appear as breaks in the line; titles use the feature name.
Feature heatmap chains	1600×400 elongated layout; each day is represented by a circular marker; values are mapped to a consistent low-to-high color gradient; day indices are shown beneath the markers.
Alternative organization	Feature visuals can be combined into a single composite image, but LE-Viz primarily uses interleaving so that each feature's visual evidence remains locally paired with its textual context.
Meta-narrative generation	External LLM prompting emphasizes global patterns, feature-specific trajectories, co-variation, and contextual interpretation; full-window narratives are typically 150–250 words while feature-level summaries are shorter.
VLM adaptation	LLaVA-NeXT 7B with LoRA adaptation; the appendix records rank 64, alpha 128, learning rate 10^{-4}, batch size 4, 20 epochs, and greedy decoding with a large token budget.
LLM text baselines	Llama 3.1 8B Instruct with LoRA; the appendix records rank 32, alpha 16, learning rate 10^{-5}, batch size 8, and 20 epochs.

separate prompt family, but it does sharpen the interpretation of the narrative target: it should capture stable behavioral structure rather than dataset-specific surface cues [22]. This appendix therefore records that shift as a procedural rule.

[title=Template I: Frozen-prior coherence target, breakable] Given a LE trajectory, generate a semantically grounded future-state narrative that captures stable behavioral structure rather than a narrow task-specific shortcut. The resulting narrative is used not only as a target text, but also as a coherence reference that constrains the learned representation.

B.13 Reproducibility Notes

Because prompt design is part of the method, it should be treated as reproducibility-critical rather than informal. Exact prompt wording, decoding settings, label-extraction behavior, and multimodal assembly choices all shape the effective input distribution seen by the model.

Table B.3: Suggested reproducibility items for the prompt and template pipeline.

Component	**What should be preserved**
Input template	exact wording of the prompt family, placeholders used, and dataset-specific field ordering
Generation settings	temperature, top-p, max tokens, stopping rules, and any frequency or repetition penalties
Narrative post-processing	normalization rules, trimming, sentence filtering, or structured extraction from free-form output
Label extraction	exact extraction prompt, allowed labels, and tie-breaking or fallback behavior
Missingness descriptors	descriptor-generation instruction, context window, and whether descriptor creation is model-generated or rule-assisted
Multimodal construction	ordering of text and images, image granularity, feature ordering, and visual normalization choices

B.14 Concluding Remarks

Across the dissertation, prompts serve many roles. They verbalize structured data, preserve absence, enforce contextual interpretation, generate narrative abstractions, coordinate cross-modal evidence, and define stable semantic targets. Considered in isolation, any one prompt might look like a practical interface string. Considered across the full dissertation, prompt evolution mirrors the representational evolution of the dissertation itself. That is why this appendix belongs in the formal record of the dissertation rather than in scattered implementation notes.