From Word Sequences to Behavioral Sequences: Adapting Modeling and Evaluation Paradigms for Longitudinal NLP

Anonymous ACL submission

Abstract

While NLP typically treats documents as independent and unordered samples, in longitudinal studies, this assumption rarely holds: documents are nested within authors and ordered in time, forming person-indexed, time-ordered *behavioral sequences*. Here, we demonstrate the need for and propose a longitudinal modeling and evaluation paradigm that consequently updates four parts of the NLP pipeline: (1) evaluation splits aligned to generalization over people (*cross-sectional*) and/or time (*prospective*); (2) accuracy metrics separating between-person differences from within-person dynamics; (3) sequence inputs to incorporate history by default; and (4) model internals that support different *coarseness* of latent state over histories (pooled summaries, explicit dynamics, or interaction-based models). We demonstrate the issues ensued by traditional pipeline and our proposed improvements on a dataset of 17k daily diary transcripts paired with PTSD symptom severity from 238 participants, finding that traditional document-level evaluation can yield substantially different and sometimes reversed conclusions compared to our ecologically valid modeling and evaluation. We tie our results to a broader discussion motivating a shift from word-sequence evaluation toward *behavior-sequence* paradigms for NLP.

1 Introduction

NLP typically frames prediction as mapping isolated instances of language to an outcome. However, recent work increasingly targets inherently longitudinal problems, where multiple instances of writing are accrued per person (Kumar and Carley, 2019; Sawhney et al., 2021; V Ganesan et al., 2021; Tsakalidis et al., 2022). In these regimes, documents are generated by *people* over *time*, yielding person-indexed, time-ordered behavioral sequences rather than independent instances that current standard statistical techniques are designed for. This mismatch propagates through the standard NLP pipeline: random document splits can leak person-specific signal and scramble temporal order, per-document training ignores informative history, and flattened accuracy metrics lack clear real-world generalization goals. As a result, models can appear stronger than they really are for realistic applications and conclusions about the relationship between language and outcomes can even be flipped (e.g. see Figure 1).

These violations of the independence assumption are not rare; they arise when many documents are produced or labeled by a limited set of authors (Geva et al., 2019), and they are further amplified in longitudinal data where documents are also ordered in time. In such "human-level" applications – such as predicting mental health, personality, or demographics from text – the effective sample size is not the number of *documents* in a dataset but the number of *individuals* (V Ganesan et al., 2021), and variation decomposes into between-person differences and within-person change (Curran and Bauer, 2011; Hoffman and Stawski, 2009). However traditionally used flattened, document-level metrics conflate these components, rewarding models that memorize stable person idiosyncrasies and obscuring whether they capture within-person dynamics (Hamaker et al., 2015).

These observations motivate an ontological shift in NLP pipeline: treat documents as ordered samples of *verbal behavior* emitted by individuals, not as word sequences (Boyd and Schwartz, 2021; V Ganesan et al., 2024; Soni et al., 2024). Formally, each observation is indexed by person i and time t, $(x_{i,t}, y_{i,t})$, with dependence induced by (a) person-level structures (style, baseline levels) and (b) temporal structures (autocorrelation, trends). Consequently, evaluation metrics should distinguish whether performance is driven by generalization to stable person-level differences, or dynamic within-person changes.

When documents are nested within individuals and ordered in time, evaluation must specify what is held out (people, time, or both), since each choice tests a different generalization problem. We therefore separate two axes: (1) *cross-sectional generalization*: does a model generalize to unseen people? and (2) *prospective generalization*: does a model generalize forward in time for the same people? These targets are not interchangeable, and collapsing them into a single document-level evaluation can yield misleading inferences about model usefulness.

Here, we introduce a modeling and evaluation framework for longitudinal ML/NLP. We show why such a framework is necessary when text is treated as a repeated measurement of a person over time. Rather than optimizing a single leaderboard endpoint, we ask what a model's performance can legitimately claim under different deployment-relevant generalization targets. We demonstrate these points on a rich longitudinal dataset, showing how standard NLP/ML evaluation choices can mischaracterize performance when documents are treated as independent examples.

We make five contributions, organized from problem to implications: (1) Demonstrating the problem: We show that traditional document-level evaluation can yield substantially different conclusions than evaluations tied to ecologically valid, real-world use-cases. (2) Developing an evaluation framework: We operationalize evaluation splits that target generalization across *people* (cross-sectional) and across *time* (prospective), and show that these settings support qualitatively different inferences. (3) Demonstrate need for metric decomposition: We propose reporting both *between-person* and *within-person* variants of standard metrics to disentangle person-level and temporal performance. (4) Show the need to adapt modeling: Incorporating temporal context improves performance under these changes and brings out longitudinal modeling challenges. (5) Broader implications: We connect these findings to a shift from word-sequence evaluation toward *behavior-sequence* paradigms for longitudinal NLP.

2 Dataset

We use the PTSD-STOP dataset (Ringwald et al., 2025) comprising dense longitudinal monitoring of PTSD symptoms with multimodal daily measurements. Participants were recruited through Stony Brook's World Trade Center Health Program[1] and completed a daily protocol for up to 90 days. Each day, participants (i) recorded a brief video diary on a personal smart device and (ii) completed a self-report PTSD symptom questionnaire (PCL; Ruggero et al., 2021). Participants received $5 per completed daily entry (video + survey), for up to $450 across the study.

	Total Data	Cross-Sectional	Prospective	Cross-Sectional & Prospective
# Participants	238	48	190	48
# Documents	17,051	2,290	4,541	1,147
Avg Docs per Person	71.6	47.7	23.9	23.9

Table 1: **Dataset statistics overall and within each evaluation regime.** Matrix icons depict how instances are partitioned for the split of interest: yellow cells indicate training person-day instances and green cells indicate the corresponding test instances (gray denotes unused). Statistics corresponds to the green region of respective column.

Daily diary videos and transcripts. In the video diaries, participants described their daily experiences and emotions in response to prompts (e.g., "*tell me about the worst part of your day*," "*describe when you felt most happy today*")[2]. We transcribe videos using the Whisper automatic speech recognition model (Radford et al., 2023) and treat each transcript as a document. The average diary length is 4.52 minutes (SD=2.12) and contains 646.6 words on average (SD=358.6).

Outcome variable. Our primary outcome is the participant's daily PTSD symptom severity score from the PCL questionnaire, denoted y_{it} for person i on day t. In this dataset, the PCL score ranges from 1 (low severity) to 5 (high severity).

Analytic sample and coverage. After filtering to participants with sufficiently dense longitudinal coverage (at least 50% days with available observations for a maximum of 90 days) and aligning diaries with outcome, our final analytic sample contains 238 participants with 17,051 documents (mean=71.6 documents/person). Table 1 summarizes dataset counts overall and within the evaluation subsets used throughout the paper.

Participant characteristics. Participants were predominantly White (60%; 4% Black/African American, 6% other/multi-racial, 29% not recorded/refused) and male (84%), with an average

[1] www.stonybrookmedicine.edu/WTC

[2] refer §A for the list of prompted questions

age of 59.87 years (range 42–79). We refer readers to Ringwald et al. (2025) for additional recruitment, protocol, and cohort details.

Tasks. We study two language-to-symptom prediction settings. For the longitudinal evaluation paradigm (§3), we consider a same-day *nowcasting* task: predicting daily PTSD severity $y_{i,t}$ from same-day language $x_{i,t}$. For the longitudinal modeling analyses (§4), we consider *one-day-ahead forecasting*: predicting next-day severity $y_{i,t+1}$ from prior-day language $x_{i,t}$.

Data preprocessing steps used for creating the datasets analyzed for the evaluation paradigm (§3) is detailed in Appendix §B.1 and that used for longitudinal modeling (§4) is detailed in Appendix §B.2.

Illustration of splits. For ease of reading, each table/figure is accompanied by a small matrix icon indicating the evaluation regime at the top. Rows correspond to people i and columns to days t; each cell represents an observed instance (e.g., $(x_{i,t}, y_{i,t})$ for nowcasting, or $(x_{i,t}, y_{i,t+1})$ for forecasting). Yellow cells denote training instances, green cells denote test instances for the evaluation split of interest, and gray cells denote instances not used for that split. Cross-sectional evaluation splits by people ($i \in \mathcal{I}_{\text{train}}$ vs. $i \in \mathcal{I}_{\text{test}}$), prospective evaluation splits by time ($t \leq \tau$ vs. $t > \tau$), and cross-sectional & prospective splits by both (train: $i \in \mathcal{I}_{\text{train}}$, $t \leq \tau$; test: $i \in \mathcal{I}_{\text{test}}$, $t > \tau$).

Missingness handling. When language is missing on a given day, we impute $x_{i,t}$ by carrying forward the most recent available language observation for that participant (last observation carried forward). In contrast, we do not impute outcomes: days with missing PCL scores ($y_{i,t}$) are excluded from training and evaluation. Accordingly, the forecasting task uses $(x_{i,t}, y_{i,t+1})$ pairs only when $y_{i,t+1}$ is observed.

3 Evaluation: Splits and Metrics for Longitudinal NLP

Traditional evaluation yielded substantially different errors than ecologically valid settings. To demonstrate the issues ensued by traditional paradigm, we mirror common settings where many documents come from a limited number of human sources (Geva et al., 2019) by samping a subset of 20 individuals from PTSD-STOP[3]. Across evaluation regimes, the *same* document-level modeling pipeline led to sharply different, and sometimes reversed conclusions (Table 2). Using a typical document-level model[4] and a baseline that predicts the training-set mean, we find that under a *traditional* random document split the model outperformed the baseline (MAE: .520 vs .660; ΔMAE=$-$ .140). However, under a more realistic *cross-sectional* split (test on unseen people), this conclusion reversed: the model underperformed the baseline (MAE: .757 vs .619; ΔMAE=+.138; $p < .001$). Under a *prospective* split (test on future days for the same people), the model again outperformed the baseline (MAE: .455 vs .598; ΔMAE=$-$ .143; $p < .001$), but the estimated error differed substantially from the traditional split (.455 vs .520).

	Ecologically Improbable	**Ecologically Valid**	
Cross Validation Split:	**Traditional Test Set** mae	**Cross Sectional Test Set** mae	**Prospective Test Set** mae
Typical model	**.520**	.757‡	**.455**‡
Baseline: Mean of Train	.660	**.619**	.598
N docs (train / test)	1008 / 421	1008 / 421	1008 / 421
Mean (Train / Test)	1.73 / 1.78	1.81 / 1.60	1.78 / 1.67
Std Dev (Train / Test)	0.79 / 0.84	0.84 / 0.71	0.84 / 0.72

Table 2: **Traditional evaluation suffers the ecological fallacy and provides no insight about generalization over people or time dimension.** This demonstration shows that ecologically invalid train - test split can lead to substantially different accuracies from valid settings using a typical model. This also shows that typical models do not necessarily work well for ecologically valid settings. For example, it leads to poor cross-sectional generalization as existing training methods regularize at document-level and not person-level. Data was sampled from 20 random people from PTSD-STOP data spanning a maximum of 90 days. *Typical model:* RoBERTa-large with a task regression head and L2 regularization. $^{\ddagger}p < .001$ (one-sided paired t-test vs. baseline).

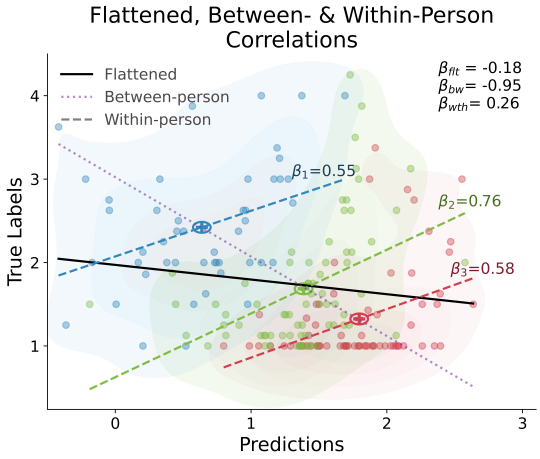
These discrepancies arise because longitudinal NLP data are person-indexed and time-ordered: random document splits can leak person identity across train and test and can induce temporal leakage (training on later observations while evaluating on earlier ones). Such an interpolation-style evaluation is typically easier than deployments that require generalization to *new people* or *future days*.

Importantly, train/test outcome means and stan-

[3] details in Appendix §B.1

[4] Fine-tuning the task-specific/regression layer of RoBERTa-large (Liu et al., 2019). Refer to Appendix §C for details.

(a) **Demonstration of Flattened metric obscuring cross-sectional and temporal evaluation of the model.**



(b) **Between- and Within-Person metrics capture how well the model generalizes over people and time.**

Metric Scope	Cross-sectional Test Set		Prospective Test Set		Cross-sectional & Prospective	
	MAE (↓)	r(↑)	MAE (↓)	r(↑)	MAE (↓)	r(↑)
Flattened	.656	.295	.434	.655	.571	.314
Between-Person	.524	.450	.230	.962	.409	.468
Within-Person	.654	.285	.426	.297	.578	.214

Figure 1: **Flattened metrics can mask whether models learn people or dynamics.** *Left:* For three illustrative users from cross-sectional test set, predictions covary within individuals, but the between-person relationship of person-level means is reversed in direction; a pooled document-level fit conflates these effects, depicted by the flattened fit. *Right:* Decomposing performance into between-person and within-person MAE/r across split regimes separates individual differences from temporal variation and clarifies what drives apparent "good" performance.

dard deviations were similar across paradigms (Table 2), and the baseline remained relatively stable, suggesting that the observed differences primarily reflect the evaluation protocol rather than large distribution shifts. Overall, Table 2 shows that the *same* modeling pipeline can appear effective, ineffective, or even harmful depending on whether evaluation targets generalization across *documents*, *people*, or *time*.

Metric Scope	Equation
Flattened	$\frac{1}{\sum_{i=1}^{n} t_i}\left(\sum_{i=1}^{n}\sum_{j=1}^{t_i} f(y_{ij},\hat{y}_{ij})\right)$
Between-Person	$\frac{1}{n}\sum_{i=1}^{n} f(\bar{y}_i,\bar{\hat{y}}_i)$
Within-Person	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{t_i}\sum_{j=1}^{t_i} f(y_{ij},\hat{y}_{ij})\right)$

Table 3: **Metric scopes for person-indexed longitudinal evaluation.** Flattened metrics pool all person-days, while between-person and within-person metrics separately evaluate person-level differences and within-person temporal variation. Here f is the per-instance metric function, and $y_{ij}, \hat{y}_{ij}$ are the observed and predicted outcomes for person i at time j.

A single pooled metric hid whether the model learned people or dynamics. Even with realistic splits, standard "flattened" metrics (computed by pooling all documents together) mixed two distinct prediction goals: (i) capturing stable differences between people (who tended to have higher scores on average) and (ii) capturing within-person change over time (day-to-day deviations). These are both important in longitudinal behavioral prediction, but a model can perform well on one while performing poorly on the other. We therefore propose reporting complementary *between-person* and *within-person* metrics (Table 3). Between-person metrics evaluate predictions after aggregating within each person (e.g., comparing each person's mean predicted score to their mean true score), while within-person metrics evaluate performance within each person first and then average across people. In simple terms, between-person evaluates how well the model captured *who was higher*, and within-person evaluates how well it captured *when someone was higher*.

Between- and within-person metrics reveal what drives "good" performance. Figure 1a illustrates why pooled evaluation could be misleading. For three example individuals from the cross-sectional test set, predictions covaried with the true labels within each person over time, yet the relationship between person-level averages differed in direction in this illustrative subset. When documents were pooled, these sources of variation were blended into a single summary that obscured which component the model captured.

Figure 1b makes this distinction explicit by

decomposing performance into between-person (individual differences) and within-person (temporal dynamics) components.[5] Across all regimes, between-person performance exceeds within-person performance. On the *cross-sectional* test set, between-person error and correlation are better than within-person (MAE: .524 vs .654; r: .450 vs .285), indicating that the model more reliably captures stable person-level differences than day-to-day variation. The gap is larger on the *prospective* test set: between-person correlation is near-perfect ($r = .962$; MAE=.230) while within-person correlation remains modest ($r = .297$; MAE=.426), implying that the strong flattened prospective correlation (flattened $r = .655$) is driven primarily by person-level signal. The combined cross-sectional & prospective setting shows the same pattern (between-person $r = .468$ vs within-person $r = .214$). Overall, reporting between- and within-person metrics alongside pooled metrics provides a clearer account of model behavior in longitudinal NLP: whether performance is driven by learning *who* differs or by tracking *how* individuals change over time.

4 Modeling: From Isolated Documents to Behavioral Sequences

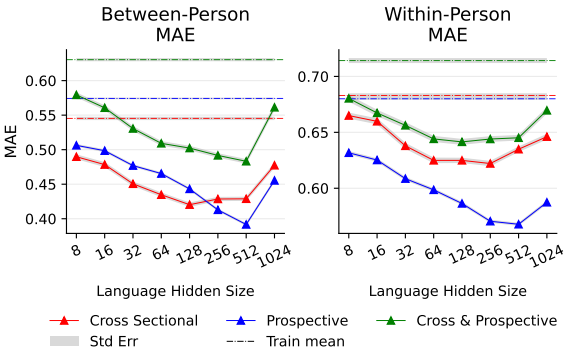


Figure 2: **Between- and within-person MAE vs. representation size.** Error shows a U-shaped trend across regimes: cross-sectional and cross-sectional & prospective perform best with small representations (64–128), whereas prospective benefits from larger size (∼512).

Representation capacity depended on the generalization target (people vs. time). We first considered the typical NLP modeling setup that maps an *isolated* document representation to an outcome, and varied the dimensionality of the language representation ("hidden size"). Prior work in human-level NLP suggested that relatively low-dimensional representations can suffice for prediction (V Ganesan et al., 2021); applying our evaluation paradigm added an important nuance. We reduced the dimensions of language using Principal Component Analysis.

As shown in Figure 2, the optimal dimensionality shifted depending on whether evaluation required generalization to *new people* (cross-sectional), to *future days* (prospective), or both (cross-sectional & prospective). In the cross-sectional and cross-sectional & prospective test sets, performance was best at relatively low dimensionality (64-128), consistent with prior qualitative finding that increasing representation size does not monotonically improve human-level prediction. In contrast, for the prospective test set, error decreased further at higher dimensionality (around 512), and this pattern held for both between-person and within-person MAE.

Across settings, prospective errors were lower than cross-sectional errors, indicating that generalizing to *unseen people* was more difficult than generalizing forward in *time* for seen people. Finally, all language-based models outperformed the "mean of train" baseline (dotted lines), and the same qualitative trends were observed under SMAPE (Figure 5).

Modeling language as behavioral *sequences* improves prediction under ecologically valid evaluation. We next tested whether treating daily diaries as a *sequence of behaviors*, rather than independent documents, improves performance. Using an autoregressive ridge model, we varied the history length h (number of prior days of language representations provided as input), where h=1 corresponds to the traditional setup.

As shown in Figure 3, incorporating longer histories reduced error across evaluation sets, with the most consistent gains in the cross-sectional split, where MAE improved monotonically with h. In the prospective split, autoregressive performance improved initially and then plateaued as history increased, while in the cross-sectional & prospective split, autoregressive performance exhibited a U-shaped relationship with h, suggesting an accuracy – complexity trade-off under the most stringent generalization setting.

Across splits, the best-performing sequence mod-

[5] We resample the data to include the third evaluation split; details are in Appendix §B.1.

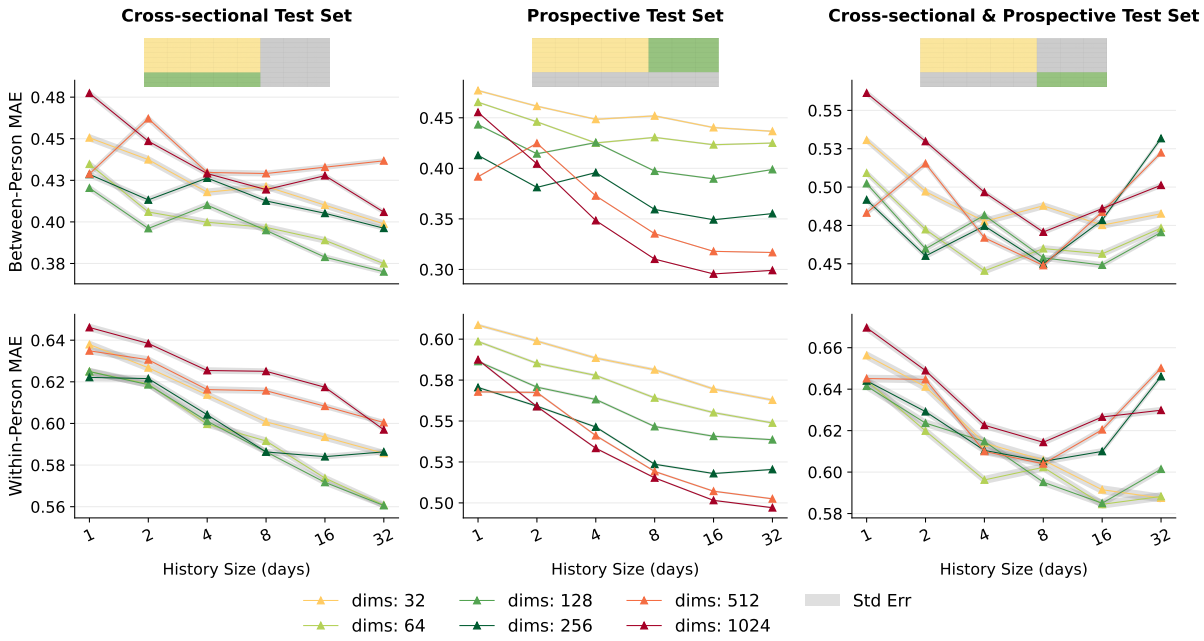


Figure 3: **Between- and within-person MAE vs. history length.** Using longer history generally improves performance, but the best history–capacity trade-off depends on the regime: prospective benefits from longer context with larger representations ($\geq$512), while cross-sectional and cross-sectional & prospective improve primarily with longer context at smaller size ($\sim$64). Across splits, the average best size is 128.

els achieved improvements that exceeded the uncertainty in our MAE estimates (standard errors were non-overlapping at the best-performing settings), indicating that the gains are robust and not driven by noise. Together, these results motivate *sequence modeling of language as the default* for longitudinal behavioral prediction.

The optimal temporal inductive bias depends on the generalization target, implying a need for *state coarseness*. To understand what aspects of history modeling drive these gains, we compared three increasingly expressive ways to map a length-h language history into a predictive state: (i) a *bagged history* (BoE) that averages the last h days' representations into a single coarse summary vector, (ii) an *autoregressive* model (AR) that explicitly parameterizes lagged dynamics over the h days, and (iii) a small *transformer* that can model content-dependent *interactions* across days (trained from scratch with a causal mask limiting attention to the past h days; Figure 4).

Strikingly, the best-performing inductive bias depended on the evaluation split. In the cross-sectional split, AR performed best, with the transformer close behind, suggesting that structured dynamics help while more flexible interaction modeling provides limited additional benefit. In the prospective split, the transformer clearly outperformed AR and BoE, consistent with the presence of predictive signal in temporal interactions once individual baselines are observed. In the cross-sectional & prospective split, BoE performed best, indicating that a *coarser*, more regularized state is advantageous under the most stringent generalization setting.

Overall, no single temporal inductive bias dominates; instead, models benefit from an internal structure that can accommodate different *levels of state coarseness*, ranging from pooled summaries, to explicit dynamics, to interaction-rich sequence models, depending on whether the task requires generalization across people, time, or both.

5 Discussion

The independence assumption is (implicitly) a human-source assumption. Our results show that assuming documents are independent is rarely an inert or inconsequential abstraction in NLP; it implicitly assumes the *human sources* that generate supervision (authors, annotators) are exchangeable and independent. Because supervision comes from a finite set of people, models can exploit source-specific regularities and fail to generalize to unseen sources, underscoring the importance of documenting dataset provenance (Geva et al., 2019; Bender

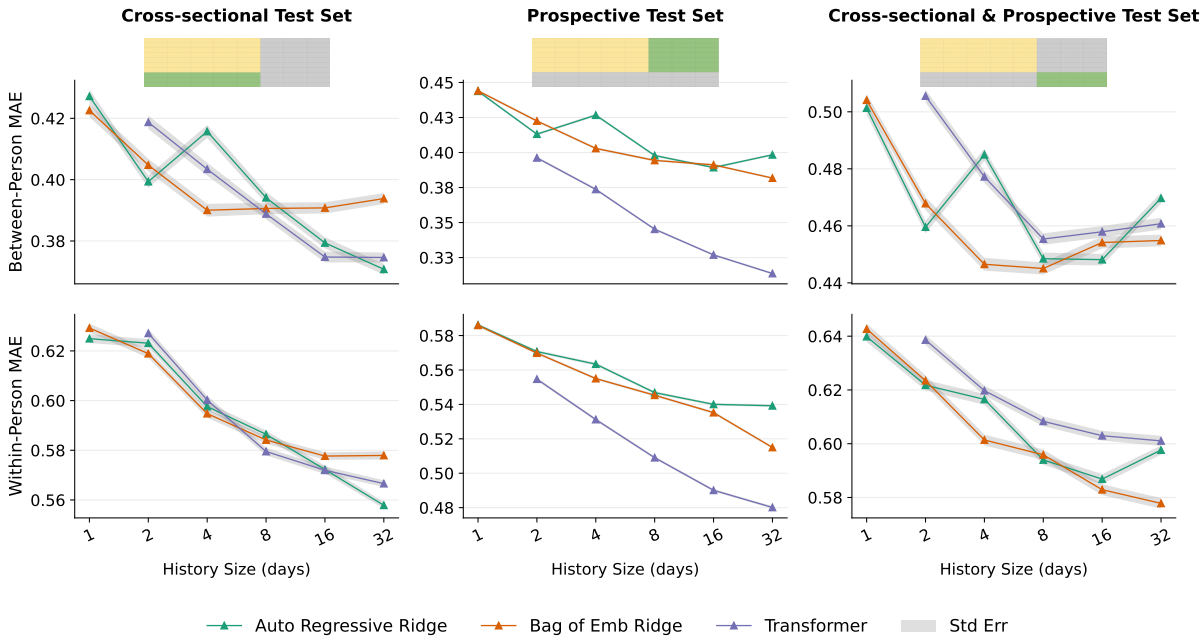


Figure 4: **AR vs. BoE vs. Transformer across history length (128 dims per day).** Between-person (top) and within-person (bottom) MAE as a function of history h. AR performs best for cross-sectional generalization, while modeling temporal interactions (Transformer) yields the largest gains for prospective generalization but performs worst for cross-sectional & prospective.

and Friedman, 2018). This aligns with human-centered NLP that treats language as *verbal behavior* produced by people in context (Boyd and Schwartz, 2021; Soni et al., 2024). Our evaluation framework operationalizes this perspective by making the intended generalization target explicit: *new documents from seen people*, *new people*, and/or *future time*.

Model evaluations lose meaning and value when they lack veridicality Our cross-sectional and prospective splits instantiate evaluation under distribution shift: performance depends on whether test-time data share the same sources (people) and time period as training, and practical deployments often restrict this overlap (e.g., predicting for *unseen individuals* or *future days*). This framing aligns with in-the-wild robustness work and evidence that models can learn "shortcuts" that look strong under convenient test sets but break under shifted conditions (Koh et al., 2021; Geirhos et al., 2020; Rosenblatt et al., 2024). In longitudinal NLP, temporally aware evaluation has been motivated as necessary for understanding behavioral dynamics rather than static aggregates (Matero and Schwartz, 2020; Tsakalidis et al., 2022). Our findings reinforce this methodological point for NLP: *evaluation design is not merely an implementation choice; it changes the scientific claim a paper is able to support.*

Metrics should reflect whether we care about *people*, *time*, or both. When documents are repeated measures of individuals, a single pooled (flattened) score can misalign with the intended behavioral claim by mixing stable between-person differences with within-person change (cf. ecological correlations Robinson, 1950). Reporting *between-person* and *within-person* variants of standard metrics makes these targets explicit and helps diagnose whether a model primarily recovers stable baselines versus tracks day-to-day variation. This mirrors measurement practice in psychology and longitudinal designs, where separating within- and between-person signal is foundational to interpretation (Shiffman et al., 2008). For NLP readers, the practical recommendation is lightweight: keep standard pooled metrics for comparability, but add target-aligned decompositions when the data are person-indexed over time.

Implications for modeling and dataset construction. Treating documents as behavioral sequences surfaces modeling choices that independence evaluation can hide, including conditioning on temporal context to model states and trajectories rather than isolated utterances (Matero et al., 2021; Soni et al., 2022). Although human-

centered shared tasks increasingly highlight longitudinal structure and temporally sensitive evaluation, these practices have not yet become default in mainstream NLP benchmarks (Tsakalidis et al., 2022; Matero and Schwartz, 2020). We therefore recommend a reporting norm for datasets with finite human sources or repeated measures: describe source structure, justify the generalization target, and align splits and metrics to that target (Bender and Friedman, 2018). Taken together, these steps move longitudinal NLP toward a methodology where conclusions are about *people and time*; not only about fitting collections of documents.

6 Related Work

Target-aligned evaluation for longitudinal NLP. A growing body of work recognizes that many applied NLP settings require temporally sensitive evaluation, especially in mental health and affective modeling (Matero and Schwartz, 2020; Tsakalidis et al., 2022). However, longitudinal evaluations often operationalize only one axis at a time (e.g., within-user temporal change or user-level holdout), making it hard to interpret what kind of generalization a reported score actually reflects. Our work complements this line by explicitly separating *cross-sectional*, *prospective*, and *cross-sectional & prospective* regimes, so evaluation claims are aligned with deployment targets.

Between-person vs. within-person variation. In "human-level" prediction, several studies emphasize that the effective sample size is the number of *people* rather than *documents*, and that performance is strongly shaped by person-level variation (V Ganesan et al., 2021). Conversely, longitudinal NLP work has highlighted within-person change as a core target(Tsakalidis et al., 2022). To our knowledge, prior work rarely unifies these two sources of variation into a single evaluation framework; our split regimes and between-/within-person metrics are designed to jointly diagnose generalization across people *and* across time.

Source-specific leakage goes beyond coarse human attributes. Related concerns appear in work on non-i.i.d. supervision, showing that models can exploit source-specific regularities and fail to generalize when the human source changes. For example, Geva et al. (2019) demonstrate that models can leverage annotator identifiers and that performance drops under annotator-disjoint splits, motivating source-aware evaluation. More recently, Orlikowski et al. (2023) show that even when sociodemographic attributes correlate with label variation, modeling those attributes does not recover individual annotator behavior, highlighting that leakage cannot be reduced to coarse metadata; our setting instantiates an analogous issue for *authors* with the added complication of temporal ordering.

Modeling people over time. A parallel line of work adapts models to individuals via user representations or personalization, and human-language modeling arguments similarly treat individuals as generators of text with states and traits (Lynn et al., 2017; Soni et al., 2022). In longitudinal prediction tasks, autoregressive and sequence-aware models have been used to exploit temporal context (Matero and Schwartz, 2020). Our contribution is complementary: we show that making the evaluation target explicit changes which modeling choices look favorable, motivating sequence inputs by default and model internals that support different coarseness of state representations depending on the regime.

7 Conclusion

We argue that traditional NLP evaluation practice implicitly assumes independence between documents, even when datasets are generated by a finite set of humans and contain repeated measures over time. Using a temporally dense daily-diary dataset, we showed that this random document-level splits can yield substantially different, and sometimes reversed conclusions relative to ecologically valid evaluation targets. To make evaluation claims explicit, we operationalized split paradigms that separately test generalization to unseen people (cross-sectional) and to future days (prospective).

We further showed that standard pooled metrics can hide what a model actually learned in longitudinal settings, and that reporting between-person and within-person metrics clarifies whether performance is driven by stable individual differences or sensitivity to day-to-day change. Finally, we connected evaluation to modeling: once the target is specified, incorporating temporal context becomes a principled modeling choice, and its benefits (and failure modes) become measurable. Overall, our goal is not a benchmark race on a single endpoint, but a practical methodological shift toward longitudinal NLP as modeling *people and time*, with evaluation protocols that support the scientific and deployment claims we want to make.

Limitations

Scope of methods and models studied. This paper is a methodology paper rather than a benchmark race, and our experiments intentionally focus on simple, transparent model classes (ridge regression and autoregressive variants) to isolate evaluation effects. As a result, we do not claim that the specific optimal representation dimensionalities or history lengths we observed will transfer unchanged to other datasets, tasks, or model families. We leave such line of work (V Ganesan et al., 2021; Singh et al., 2025) to future works. More expressive sequential architectures, additional modalities, or alternative training objectives may change absolute performance (Matero et al., 2021; Rao et al., 2025), but the core risk we highlight—that convenient independence-assumed evaluation can reward reliance on source-specific regularities and fail under shifted regimes—has been observed broadly in ML under distribution shift and shortcut learning (Koh et al., 2021; Geirhos et al., 2020). Finally, we study one clinical endpoint (daily PCL) as a concrete case; extending this framework to other outcomes (e.g., functional impairment, treatment response) and to other longitudinal NLP settings is an important direction for future work.

Dataset- and measurement-specific constraints. Our empirical demonstrations are grounded in a single intensive longitudinal cohort (PTSD-STOP) with daily self-reports and diary entries, which enables the methodological analyses we pursue but also constrains external validity. First, the cohort's demographics and recruitment context (a clinically monitored trauma-exposed population) may not represent other clinical groups or the general population, and prior work has cautioned that clinical NLP systems can inherit and amplify such representational skews if generalization targets and subgroup performance are not made explicit (Bear Don't Walk et al., 2022; Obermeyer et al., 2019). Second, intensive longitudinal data are rarely missing-at-random: adherence varies over time and across individuals, and missingness mechanisms can bias both modeling and evaluation if not carefully characterized (Shiffman et al., 2008; Stone et al., 2023; Heron et al., 2017). Third, we rely on automatic speech recognition to obtain transcripts; transcription errors and diarization artifacts can introduce additional noise that may differentially affect within-person versus between-person signals, and should be treated as part of the measurement process rather than ignored (Shiffman et al., 2008; Stone et al., 2023).

Ethical Considerations

Sensitive human-subject data and governance. This work analyzes longitudinal materials linked to mental health symptom reports, which are inherently sensitive and may contain personally identifying information (even after transcription). Accordingly, analyses should be conducted under appropriate human-subject oversight and data-use agreements consistent with established principles for research with human participants (U.S. Department of Health & Human Services, 1979). Because de-identification is not a guarantee of anonymity, especially for rich narrative data, we follow the conservative stance that access controls and minimization of shared artifacts (e.g., avoiding release of raw text/video) are often necessary for participant protection (U.S. Department of Health & Human Services, 2025; Ayers et al., 2018). We also view dataset provenance and documentation as part of ethical reporting in NLP: clearly stating who produced the data (sources), under what conditions, and what the intended generalization target is helps prevent misuse and misinterpretation of results (Bender and Friedman, 2018; Gebru et al., 2021).

Risks of downstream use in mental health contexts. Models that predict mental health–related outcomes can be misapplied in ways that harm individuals: erroneous inferences may contribute to stigma, inappropriate monitoring, or decisions made without clinical context. Prior work in digital mental health NLP emphasizes that prediction should not be conflated with diagnosis, that construct validity is difficult to establish from language traces alone, and that failure modes are often context-dependent (Chancellor and De Choudhury, 2020; Ernala et al., 2019). More broadly, the field has called for explicit ethics disclosures and clearer articulation of intended use, beneficiaries, and plausible harms in mental health prediction research (Ajmani et al., 2023; Chancellor et al., 2019). Consistent with this, our framing and experiments are aimed at improving *methodology* (what we can claim from evaluation), not at promoting clinical deployment.

Re-identification and leakage considerations. Longitudinal diaries are especially vulnerable to

re-identification because they accumulate unique life details over time. Even when direct identifiers are removed, verbatim excerpts can enable reverse identification, and model artifacts can inadvertently memorize or surface sensitive details (Ayers et al., 2018; Chancellor et al., 2019). For this reason, releasing raw text, audio, or video should be treated as a high-risk action; safer alternatives include releasing code, evaluation splits, and aggregate statistics, or providing controlled access mechanisms where appropriate (U.S. Department of Health & Human Services, 2025; Gebru et al., 2021).

Fairness, subgroup validity, and clinical safety. Clinical NLP systems can encode uneven performance across demographic or clinical subgroups, potentially exacerbating existing inequities if used in practice (Bear Don't Walk et al., 2022; Obermeyer et al., 2019). A practical implication for longitudinal NLP is that "ecologically valid" splits should be complemented with subgroup-aware reporting whenever sample sizes permit, and that claims should be scoped to the population represented by the data (Mitchell et al., 2019; Bender and Friedman, 2018). Finally, mental health applications require particular caution: as emphasized by the CLPsych community, ethical deployment demands careful consideration of consent, stakeholder involvement, and clinical safety boundaries beyond standard ML evaluation (Orr et al., 2022).

Sensitive human-subject data and governance. This work analyzes longitudinal materials linked to mental health symptom reports, which are inherently sensitive and may contain personally identifying information (even after transcription). Accordingly, analyses should be conducted under appropriate human-subject oversight and data-use agreements consistent with established principles for research with human participants (U.S. Department of Health & Human Services, 1979). The study protocol was reviewed and approved by the relevant Institutional Review Board (IRB), and all researchers followed institutional Human Subjects Research guidelines for secure data handling (e.g., restricted access, secured storage/compute, and minimization of data movement and shared artifacts). Because de-identification is not a guarantee of anonymity, especially for rich narrative data, we follow the conservative stance that access controls and minimization of shared artifacts (e.g., avoiding release of raw text/video) are often necessary for participant protection (U.S. Department of Health & Human Services, 2025; Ayers et al., 2018).

Use of AI assistance. We used AI assistants to support drafting (e.g., paraphrasing for clarity) and coding/formatting during manuscript preparation. All generated text and code were reviewed and verified by the authors, and all co-authors participated in editing and auditing to ensure accuracy, appropriate attribution, and consistency with the study protocol and results.

References

Leah Hope Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A systematic review of ethics disclosures in predictive mental health research. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1311–1323.

John W. Ayers, Theodore L. Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *npj Digital Medicine*, 1:30.

Oliver J. Bear Don't Walk, H. Reyes Nieva, S. S. J. Lee, and Noémie Elhadad. 2022. A scoping review of ethics considerations in clinical natural language processing. *JAMIA Open*, 5(2):ooac039.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Ryan L. Boyd and H. Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 79–88.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):43.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Patrick J Curran and Daniel J Bauer. 2011. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual review of psychology*, 62(1):583–619.

Sindhu K. Ernala, Michael L. Birnbaum, Asra F. Rizvi, William A. Sterling, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–16.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. In *Communications of the ACM*, volume 64, pages 86–92.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Ellen L Hamaker, Rebecca M Kuiper, and Raoul PPP Grasman. 2015. A critique of the cross-lagged panel model. *Psychological methods*, 20(1):102.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kristin E. Heron, R. Scott Everhart, Susan M. McHale, and Joshua M. Smyth. 2017. Ecological momentary assessment (ema) of youth behavior: A systematic review and recommendations. *Journal of Pediatric Psychology*, 42(10):1087–1107.

Lesa Hoffman and Robert S Stawski. 2009. Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in human development*, 6(2-3):97–120.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. 2023. The impact of positional encoding on length generalization in transformers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, 12(1):3918.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Anand Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Li, Irena Jia, et al. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Sumeet Kumar and Kathleen Carley. 2019. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5047–5058, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Matero and H. Andrew Schwartz. 2020. Autoregressive affective language forecasting: A self-supervised task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2913–2923, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2021. MeLT: Message-level transformer with masked document representations as pre-training for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2959–2966, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In

Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT), pages 220–229.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Martin Orr, Sophie Haffner, Nemanja Vaci, David Jurgens, Koen Tromp, Konstantin Wiemann, et al. 2022. The ethical role of computational linguistics in digital mental health. In *Proceedings of the 8th Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Rajath Rao, Adithya V Ganesan, Oscar Kjell, Jonah Luby, Akshay Raghavan, Scott Feltman, Whitney Ringwald, Ryan L. Boyd, Benjamin Luft, Camilo Ruggero, Neville Ryant, Roman Kotov, and H. Andrew Schwartz. 2025. WhiSPA: Semantically and psychologically aligned whisper with self-supervised contrastive and student-teacher learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22529–22544, Vienna, Austria. Association for Computational Linguistics.

Whitney R. Ringwald, Scott Feltman, H. Andrew Schwartz, Dimitris Samaras, Christopher Khudari, Benjamin J. Luft, and Roman Kotov. 2025. Day-to-day dynamics of facial emotion expressions in posttraumatic stress disorder. *Journal of Affective Disorders*, 380:331–339.

WS Robinson. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357.

Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. 2024. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15(1):1829.

Camilo J Ruggero, Keke Schuler, Monika A Waszczuk, Jennifer L Callahan, Ateka A Contractor, Charles B Bennett, Benjamin J Luft, and Roman Kotov. 2021. Posttraumatic stress disorder in daily life among world trade center responders: Temporal symptom cascades. *Journal of psychiatric research*, 138:240–245.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190, Online. Association for Computational Linguistics.

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60, Copenhagen, Denmark. Association for Computational Linguistics.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4(1):1–32.

Khushboo Singh, Vasudha Varadarajan, Adithya V. Ganesan, August Håkan Nilsson, Nikita Soni, Syeda Mahwish, Pranav Chitale, Ryan L. Boyd, Lyle Ungar, Richard N. Rosenthal, and H. Andrew Schwartz. 2025. Systematic evaluation of auto-encoding and large language model representations for capturing author states and traits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18955–18973, Vienna, Austria. Association for Computational Linguistics.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, Mexico City, Mexico. Association for Computational Linguistics.

Arthur A. Stone, Saul Shiffman, Audie A. Atienza, and Linda Nebeling. 2023. Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, 19:1–27.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

U.S. Department of Health & Human Services. 1979. The belmont report: Ethical principles and guidelines for the protection of human subjects of research. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/. Last reviewed Aug 26, 2024.

U.S. Department of Health & Human Services. 2025. Guidance regarding methods for de-identification of protected health information in accordance with the HIPAA privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Last reviewed Feb 3, 2025.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Andrew Schwartz. 2023. Systematic evaluation of GPT-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Adithya V Ganesan, Siddharth Mangalik, Vasudha Varadarajan, Nikita Soni, Swanie Juhng, João Sedoc, H. Andrew Schwartz, Salvatore Giorgi, and Ryan L Boyd. 2024. From text to context: Contextualizing language with humans, groups, and communities for socially aware NLP. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 26–33, Mexico City, Mexico. Association for Computational Linguistics.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

A PTSD-STOP Daily Diary Study

Open Ended Questionnaire Each day, participants completed a brief video diary consisting of 13 open-ended prompts about their day. The camera began recording automatically and the recording uploaded automatically upon submission. Participants were instructed to answer each prompt out loud while facing the camera, to expand on their responses (rather than giving short answers), and to avoid reading the questions aloud. The submit button appeared after 3 minutes, with up to 10 minutes available per prompt. The 13 prompts were:

1. Tell me about the best part of your day.
2. Tell me about the worst part of your day.
3. Describe when you felt most sad today.
4. Describe when you felt most scared or nervous today.
5. Describe when you felt most annoyed today.
6. Describe when you felt most happy today.
7. How did you feel physically today? Did you have any pain, discomfort, or other physical symptoms? Please elaborate.
8. How did you get along with others today? Please elaborate.
9. Did you have any unwanted, disturbing memories of a past stressful experience? Tell me about this.
10. Today, did you avoid anything because it would have made you uncomfortable? Tell me more about what you did and why.
11. Did you feel on guard today? What made you feel this way?
12. Describe anything that cheered you up today. How did it go?
13. What happened today that you can feel thankful for? Tell me about this.

Rating Scales In addition to the open-ended diary, participants completed daily self-report rating scales capturing (i) post-stressor/PTSD-related experiences and overall stress that day, and (ii) exposure to common daily stressors. Rating scales were administered in two parts.

Part 1 (Symptom and stress ratings). Participants rated the following items everyday, keeping their most stressful event in mind. Response options were: *Not at all, A little bit, Moderately, Quite a bit, Extremely*, and *Skip*.

1. Today, I had repeated, disturbing, and unwanted memories of the stressful experience.
2. Today, I felt very upset because something reminded me of the stressful experience.
3. Today, I avoided memories, thoughts, or feelings related to the stressful experience.
4. Today, I avoided external reminders of the stressful experience (for example, people, places, conversations, activities, objects, or situations).
5. Today, I felt distant or cut off from other people.
6. Today, I had strong negative feelings such as fear, horror, anger, guilt, or shame.
7. Today, I felt jumpy or easily startled.
8. Today, I was "superalert" or watchful or on guard.
9. Overall, how stressed did you feel today?

Part 2 (Daily stressor checklist). Participants then indicated which of the following troublesome or stressful events occurred everyday (check all that apply):

1. Had tension or argument with spouse, partner or close family
2. Had tension or argument with others (e.g., coworker, friend, etc)
3. A lot of demands at home
4. A lot of demands at job
5. A lot of demands made by family
6. Caring for a sick family member
7. Problems with transportation
8. Financial or money problem(s)
9. Health-related event(s)
10. Other troublesome things happened to me
11. No troublesome or stressful things happened to me today

B Data Pre-processing

B.1 Longitudinal Evaluation

The violations of independence assumption arise whenever many documents are produced or labeled by a limited set of humans, and they are amplified in longitudinal data where documents are also ordered in time. Such dependencies are more often the rule than the exception. Even when an NLP task may appear on its surface to be "about the text itself" (e.g., classification, translation, summarization), supervision originates from a finite set of human sources: annotators create labels, and translators/summarizers create references. In practice, the "human set" can be surprisingly small relative to the number of documents, creating a deeply shared variance structure across instances.

For example, crowdsourced datasets often show extreme annotator imbalances: in the NLU datasets analyzed by Geva et al. (2019), MNLI contains 402k examples from 380 annotators and OpenBookQA contains 5k examples from 84 annotators, with the most prolific OpenBookQA annotator contributing 24% of all examples. Likewise, widely used summarization corpora contain reference summaries written within a limited set of institutions: Newsroom provides 1.3M article – summary pairs written by authors and editors across 38 news publications (Grusky et al., 2018).

Demonstration of ecological fallacy in traditional evaluation. Analyses for Table 2 and Figure 1a use the dataset constructed as follows. To mirror common settings where many documents come from a limited number of human sources (Geva et al., 2019), we form a controlled subset of 20 individuals from PTSD-STOP.

We first filtered to individuals with non-trivial outcome variation, requiring at least 0.01 standard deviation in PCL over the full 90-day period, as well as within the first 60 days and last 30 days. We then sorted individuals by their mean PCL (averaged over days), binned them into 10 strata, and sampled two individuals uniformly at random from each stratum to obtain a balanced cohort of 20 people.

We split this cohort into train/test in three ways: (a) a document-level random split (70/30), which is ecologically implausible for person-indexed data; (b) a cross-sectional split (70/30) by person, which holds out individuals; and (c) a prospective split (70/30) by time, using a temporal cutoff at day

$\tau = 63$ (70% of 90 days). To match train/test sizes across these three settings, we randomly masked a small number of person-day instances.

Between- and within-person metrics isolate what drives performance. Analyses for Figure 1b use the same 20-person cohort as above, but evaluate one fixed model across multiple test regimes. We set the prospective temporal cutoff to a 67/33 split (i.e., $\tau \approx 60$) to define the training window, and we additionally construct a cross-sectional & prospective test set that requires generalization over both people and time.

Concretely, we train a single model on the training region (yellow cells in the split illustration) and evaluate that same model under three regimes: cross-sectional (held-out people), prospective (held-out future days), and cross-sectional & prospective (held-out people at held-out future days). This design ensures differences in performance reflect the evaluation regime rather than training different models.

B.2 Longitudinal Modeling

For all results in §4, we use the full analytic sample of 238 participants (Table 1). For cross-sectional evaluation, participants are stratified into train/test (80/20) based on their mean PCL over the study period. For prospective evaluation, we use a fixed temporal cutoff at day $\tau = 60$ for training, with later days reserved for testing.

C Model Training

C.1 Longitudinal Evaluation

Typical model. Our *typical* model is fine-tuning of the task-specific ridge regression layer of a RoBERTa-large encoder (Liu et al., 2019). We use this term because fine-tuned encoder models remain a strong and widely used baseline for psychological and behavioral prediction from language (Kjell et al., 2022), and recent evidence suggests that instruction-tuned LLMs do not consistently outperform fine-tuning smaller encoder models for behavioral prediction and psychological measurement (Singh et al., 2025; V Ganesan et al., 2023; Choi et al., 2023).

In practice, we train ridge regression models on transcript representations extracted from the second-to-last layer of RoBERTa-large using DLATK (Schwartz et al., 2017). Language dimensions were reduced using Principal Component Analysis whenever a hidden dimension size of le 1024 was used. The ridge penalty (L2 regularization strength) is selected from $\{10^i : i \in [-2, 5]\}$ based on performance on a development set. To avoid leakage, the development set is split from the training data using the same evaluation regime as the corresponding test set (cross-sectional, prospective, or cross-sectional & prospective).

C.2 Longitudinal Modeling

Autoregressive ridge models. Autoregressive ridge models take as input a length-h history of daily language representations and predict the target outcome (nowcasting or one-day-ahead forecasting, depending on the experiment). These models are parameterized by history length h and per-day representation size d, yielding an input dimension of $h \cdot d$. Training and hyperparameter selection follow the same procedure as above (ridge regression with the L2 penalty chosen on a regime-matched development split).

Bag-of-embeddings models. Bag-of-embeddings (BoE) models aggregate a length-h history by averaging the h daily representations into a single vector, which is then used for ridge regression. As a result, BoE is parameterized by history length h and per-day representation size d, but its number of learned parameters scales only with d (since the history is pooled before prediction). We train BoE models with the same ridge procedure and select the L2 penalty on a regime-matched development split.

Transformers. Model architecture. To test whether modeling *temporal interactions* improves prediction beyond pooled summaries (BoE) and linear dynamics (AR), we trained a minimal autoregressive transformer over the history sequence. We project each per-day representation from d=128 down to d'=32 using a learned linear map, and feed the resulting sequence into a transformer encoder with 1 layer and 1 attention head (hidden size 32).

For comparability with AR and BoE, we made two design choices: (1) *No positional embeddings.* We omit positional embeddings to reduce explicit order information, isolating gains attributable to interaction modeling rather than architectural encodings of dynamics; prior work also suggests transformers can recover positional information even without explicit embeddings (Haviv et al., 2022; Kazemnejad et al., 2023). (2) *Fixed history masking.* For a given history length h, we apply a lower-

triangular attention mask and additionally prevent attention to tokens more than h days in the past by setting the corresponding attention logits to $-\infty$ prior to softmax. This ensures each transformer configuration is trained and evaluated under the same effective history constraint as the corresponding AR/BoE setting.

Hyperparameter optimization. We tuned the projection size d', L2 weight decay, attention dropout, output-layer dropout, and learning rate on a regime-matched development split. We searched projection sizes among powers of two below 128 and selected d'=32. The selected hyperparameters were: weight decay $= 1$, attention dropout $= 0.3$, output dropout $= 0.1$, and learning rate $= 10^{-3}$ with AdamW.

Additional checks. Although evaluating full-capacity transformers (e.g., larger depth/width, unmasked attention, or explicit sinusoidal/rotary positional embeddings (Vaswani et al., 2017; Su et al., 2024)) is beyond our scope, we verified that adding positional embeddings did not yield consistent improvements over our minimal configuration, and that increasing the maximum training history produced the same qualitative patterns reported in Figure 4.

D Metric Functions

We optimize models using the standard (flattened) mean squared error, leaving alternative objectives and metric formulations for future work. For evaluation, we report between- and within-person variants of Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Pearson correlation (r). SMAPE and r are bounded and scale-invariant, ranging in $[0, 2]$ and $[-1, 1]$, respectively.

SMAPE. For predictions $\hat{y}$ and targets y, we compute

$$\mathrm{SMAPE}(y, \hat{y}) = \frac{2}{N} \sum_{n=1}^{N} \frac{|\hat{y}_n - y_n|}{|y_n| + |\hat{y}_n| + \epsilon}, \quad (1)$$

where N is the number of evaluated instances and ϵ is a small constant to avoid division by zero. Since the outcomes range between [1, 5], ϵ was set to 0.

Pearson correlation. We compute Pearson correlation between y and $\hat{y}$ as

$$r(y, \hat{y}) = \frac{\sum_{n=1}^{N} (y_n - \bar{y})\,(\hat{y}_n - \bar{\hat{y}})}{\sqrt{\sum_{n=1}^{N} (y_n - \bar{y})^2}\,\sqrt{\sum_{n=1}^{N} (\hat{y}_n - \bar{\hat{y}})^2}}, \quad (2)$$

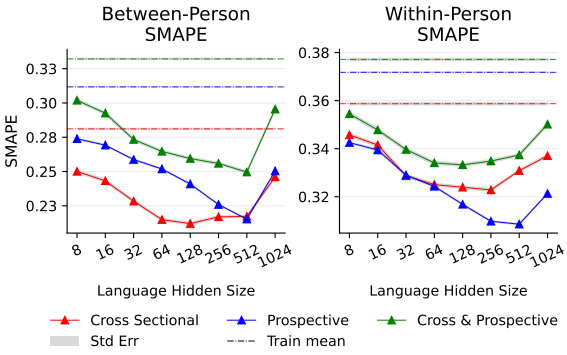


Figure 5: **Between-** ***(top)*** **and Within-Person** ***(bottom)*** **SMAPE as a function of hidden dimension size.** Forecasting performance follows a U-shaped trend as a function of hidden dimension size of language across all three evaluation sets. While a typical model requires only 64 dimensions of language for best performance on Cross-sectional and Cross-sectional & Prospective test sets, it requires 512 dimensions in Prospective evaluation set. Based on the best performance achieved in different settings, generalization to Cross-sectional & Prospective is the hardest, followed by Cross-sectional set and finally prospective set. Generalization to unseen people is harder than unseen time and within-person changes is harder than between-person differences.

where $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$ and $\bar{\hat{y}} = \frac{1}{N}\sum_{n=1}^{N} \hat{y}_n$.

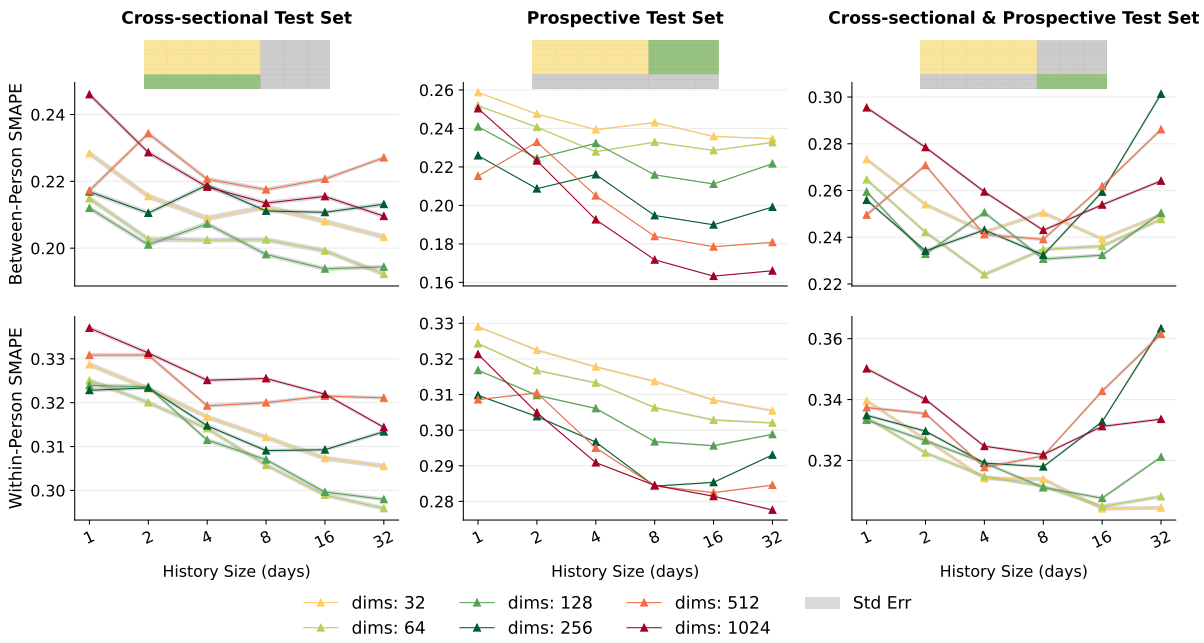
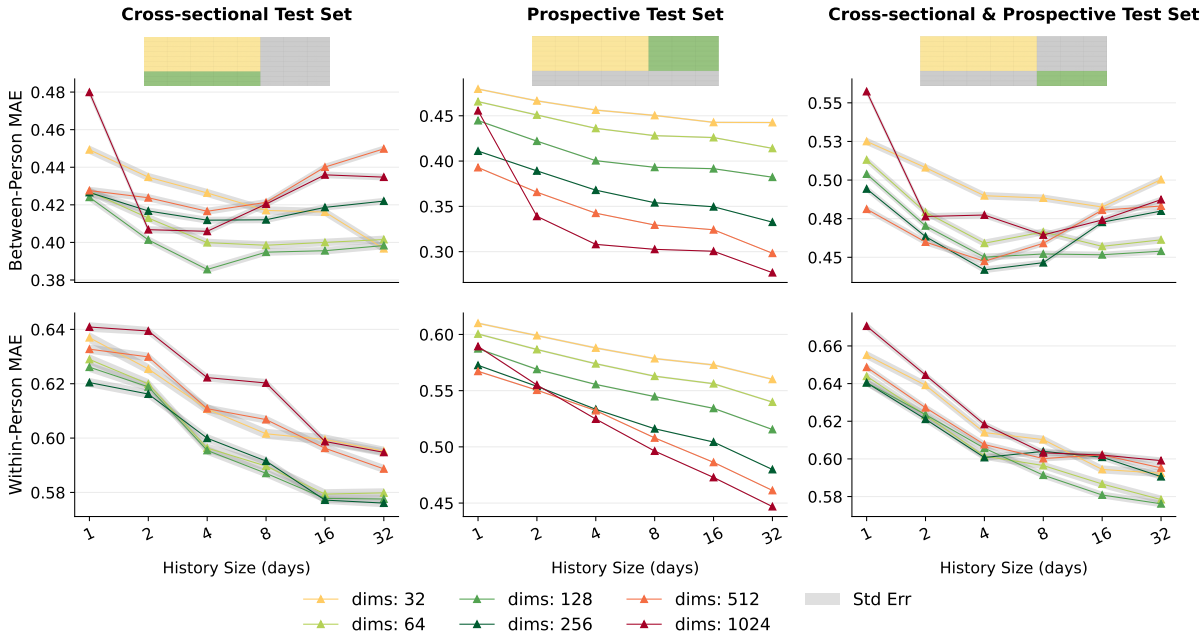


Figure 6: **Between- and Within-Person SMAPE as a function of history length for Auto Regressive Model.** Predictive Performance improves with modeling the temporal dynamics of linguistic behavior. For Prospective test set, between- and within-person performance improves with longer temporal context and higher dimensional sizes (512 and 1024). For Cross-sectional and Cross-sectional & Prospective evaluation sets, performance improves with temporal context at lower dimensions (hidden size=64).

(a) **Between-** ***(top)*** **and Within-Person** ***(bottom)*** **MAE as a function of history length for Bag of Embeddings Model.**



(b) **Between-** ***(top)*** **and Within-Person** ***(bottom)*** **SMAPE as a function of history length for Bag of Embeddings Model.**

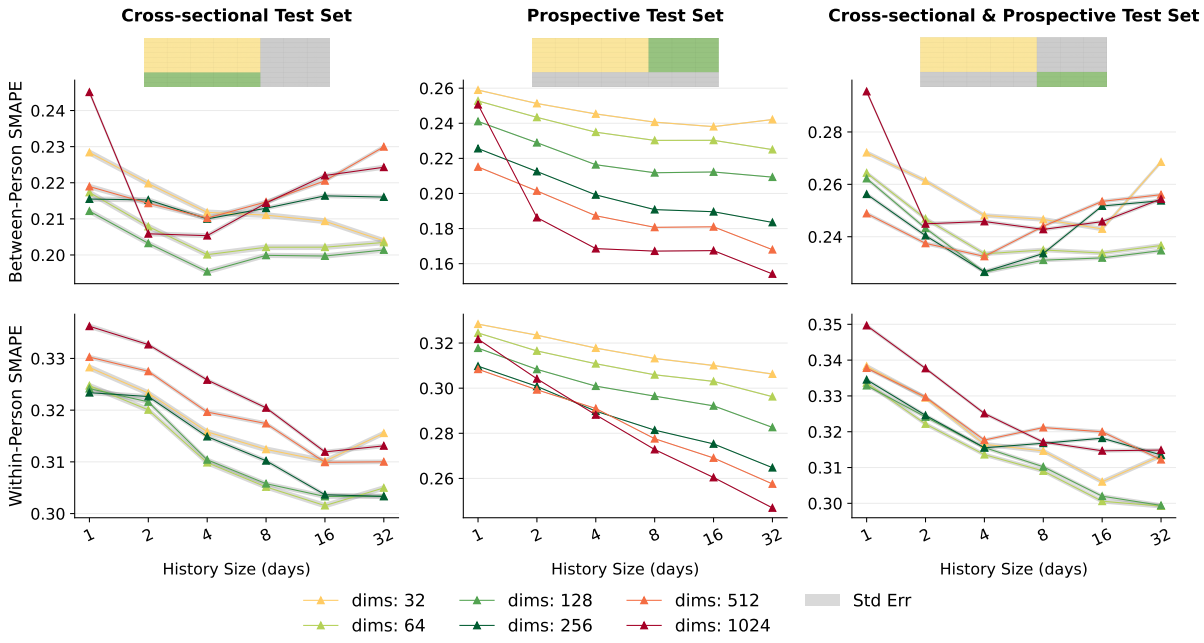


Figure 7: **Between- and Within-Person metrics as a function of history length for Bag of Embeddings Model.** Predictive Performance improves with modeling the temporal context of linguistic behavior. For Prospective test set, between- and within-person performance improves with longer temporal context and higher dimensional sizes (512 and 1024). For Cross-sectional and Cross-sectional & Prospective evaluation sets, performance improves with temporal context at lower dimensions (hidden size=64-128).

Figure 8: **Comparison of Auto Regressive, Bag of Embeddings and Transformer model using Between-** ***(top)*** **and Within-Person** ***(bottom)*** **SMAPE as a function of history length.**

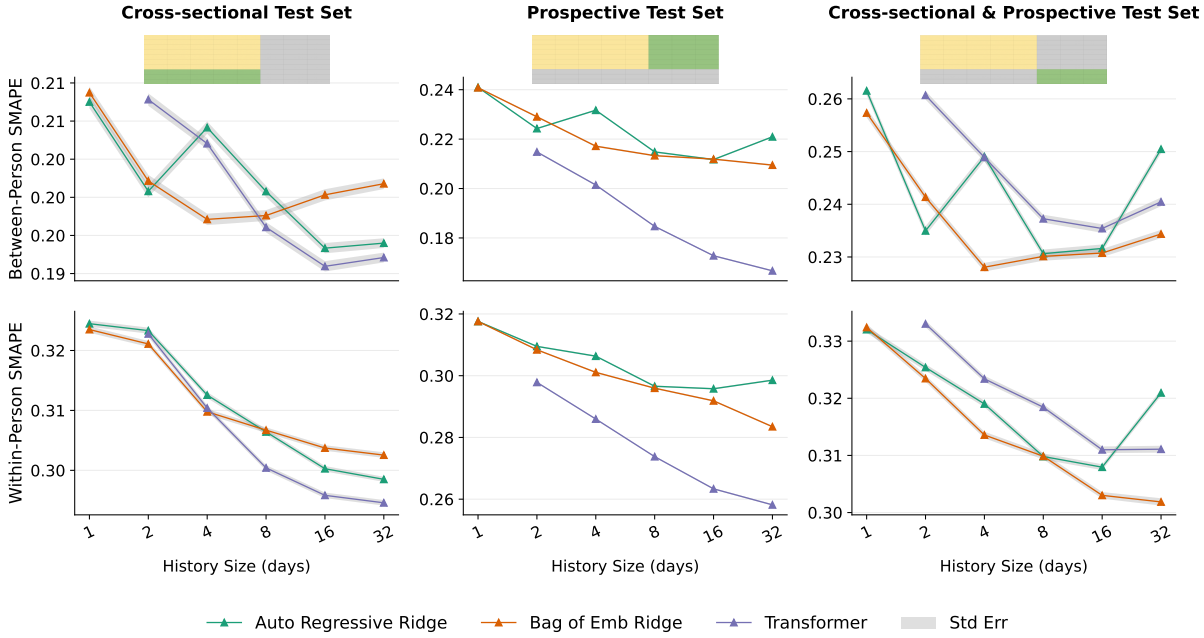


Figure 9: **AR vs. BoE vs. Transformer across history length (128 dims per day).** Between-person (top) and within-person (bottom) SMAPE as a function of history h. Modeling temporal interactions (Transformer) offers marginal improvements over modeling dynamics in cross-sectional splits, but yields the largest gains for prospective generalization. It performs worst for cross-sectional & prospective.