DYNALDI: DYNAMIC LATENT STATE ESTIMATION WITH TIME-VARYING WEAK-SIGNAL RELIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal language analyses often aggregate heterogeneous weak signals, such as lexicon scores, sentiment measures, and neural classifier outputs, into a single temporal index despite changes in measurement quality and corpus composition. We introduce **DynaLDI**, a dynamic latent-index model that treats these signals as noisy measurements of a shared longitudinal state with autoregressive dynamics, source-specific nuisance effects, and reliability-conditioned time-varying measurement variance. On 480 held-out known-truth simulated longitudinal sequences spanning ten matched and misspecified regimes, the pre-specified DynaLDI configuration reduces latent-state RMSE by 24.0% relative to equal-weight standardized aggregation and by 7.0% relative to an earlier dynamic estimator, outperforming the latter in 91.9% of paired trials. It recovers source loadings with mean cosine similarity 0.999 and time-varying precision with correlation 0.985, while nominal 95% intervals attain 94.9% coverage under matched and near-matched conditions. Applied without retuning to 1.52 million `r/depression` submissions from 2009–2022, DynaLDI reduces several associations with observable corpus drift. Real-data perturbations reveal substantially greater sensitivity to measurement-model re-identification than to conditional state inference. A separate 384-sequence known-truth benchmark shows that the resulting composite sensitivity envelope improves severe-misspecification coverage from 71.6% to 84.4%, although its posterior-variance fraction is not a calibrated local reliability score. Prespecified ecological comparisons show selective temporal convergence with CDC depressive-symptom estimates, NSDUH adult major depressive episodes, and Google Trends searches, while U.S. suicide mortality remains negatively associated. The state-space computational kernel scales approximately linearly with sequence length. The Reddit trajectory is interpreted as a relative latent state of depression-related discourse measurements, not clinical prevalence, diagnosis, or individual risk.

1 INTRODUCTION

Longitudinal language studies often construct temporal indices by standardizing and averaging heterogeneous signals such as lexicon frequencies, sentiment scores, and neural-model outputs. Although simple, this procedure implicitly treats the measurements as comparably informative and assumes that their relationship to the underlying phenomenon remains stable. These assumptions are particularly restrictive for longitudinal social-media data, where post volume, user participation, available text, moderation, deletion, and document length can change substantially over time.

We study a more general problem: *how can a shared longitudinal state be estimated from multiple weak linguistic measurements whose nuisance dependence and reliability change over time?* The measurements are *weak* because none is treated as ground truth. Lexicon scores are interpretable but narrow, sentiment tools measure broader affective properties, and neural models may be expressive while remaining sensitive to training-task and domain differences. The target is therefore the temporal component jointly supported by several imperfect measurements after observable variation in the measurement process is accounted for.

We introduce **DynaLDI** (Dynamic Latent State Estimation with Time-Varying Weak-Signal Reliability), a linear-Gaussian dynamic measurement model. Each source observes a common scalar

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

latent state through a source-specific loading and nuisance model, while measurement variance is allowed to change with observable reliability covariates. The latent state evolves autoregressively, enabling principled propagation through missing observations. Source loadings remain fixed while effective source influence changes through time-varying precision, reducing the identification ambiguity that would arise if both loadings and variances varied freely. Retrospective estimates use Rauch–Tung–Striebel smoothing, while the corresponding filtered state uses observations only through time t, conditional on globally fitted measurement parameters (Kalman, 1960; Rauch et al., 1965; Durbin & Koopman, 2012).

A second focus is the distinction between *conditional state uncertainty* and *measurement-model identification stability*. A state posterior may be narrow conditional on a fitted measurement model even when re-estimation under plausible changes in sources, temporal support, or data quality identifies a materially different measurement decomposition. We therefore supplement posterior uncertainty with perturbation-based identification and support diagnostics and evaluate these diagnostics separately on known-truth simulations. The resulting composite sensitivity envelope is treated as a model-health diagnostic rather than a Bayesian credible interval or probability of correctness.

We evaluate DynaLDI in two settings. First, a pre-specified confirmatory benchmark contains 480 held-out simulated longitudinal sequences across ten matched and misspecified data-generating regimes. DynaLDI reduces latent-state RMSE by 24.0% relative to equal-weight standardized aggregation (Equal-Z) and by 7.0% relative to an earlier dynamic estimator, while accurately recovering source loadings and time-varying measurement precision. Second, we construct a daily five-source panel from 1,516,442 `r/depression` submissions from 2009–2022 using lexical, sentiment, and transformer-derived measurements. This application is interpreted strictly as a relative state of depression-related discourse measurements, not as depression prevalence or individual clinical status.

2 Related Work

Longitudinal language and weak measurement. Longitudinal NLP has modeled temporal change through approaches including dynamic topic models (Blei & Lafferty, 2006), diachronic representations (Hamilton et al., 2016), and aggregated lexical or affective indices (Golder & Macy, 2011; Dodds et al., 2011). Weak-supervision methods similarly combine imperfect information sources rather than assuming a single ground-truth measurement (Dawid & Skene, 1979; Ratner et al., 2016; 2017). DynaLDI addresses a complementary setting in which heterogeneous linguistic measurements—for example, lexicon frequencies, sentiment scores, and neural contrast scores—are continuous longitudinal observations that may differ in orientation, nuisance sensitivity, and reliability. Rather than averaging standardized features, we model them as noisy measurements of a shared latent state whose effective source contributions may change over time.

Dynamic latent-variable models and changing measurement quality. Linear state-space and dynamic-factor models provide established tools for separating latent temporal dynamics from noisy observations (Kalman, 1960; Rauch et al., 1965; Durbin & Koopman, 2012; Stock & Watson, 2002). Dynamic factor models have also been extended to changing measurement properties and time-varying parameters (Chow et al., 2011; Poncela et al., 2021), while related multi-source work has considered jointly evolving quantities and source reliability (Li et al., 2015). Measurement-error and predictive-uncertainty literatures further emphasize that structured observation error and dataset shift can invalidate uncertainty estimates when the measurement process changes (Fuller, 1987; Kendall & Gal, 2017; Ovadia et al., 2019). DynaLDI therefore does not claim novelty for Kalman filtering, dynamic factors, heteroskedastic measurement models, or time-varying reliability individually. Its contribution is their integration into an interpretable weak-measurement formulation for longitudinal language: observable corpus conditions enter source-specific mean and precision models, semantic loadings remain fixed for identifiability, and perturbation analyses explicitly distinguish conditional state uncertainty from measurement-model re-identification sensitivity.

Mental-health language modeling. Prior work has used social-media language for depression-related classification, behavioral characterization, shared-task prediction, and longitudinal forecasting (De Choudhury et al., 2013; Coppersmith et al., 2014; 2015; Benton et al., 2017; Reece et al., 2017), with transformer models expanding the available representation families (Devlin et al., 2019;

Liu et al., 2019). Our objective is different from individual-level diagnosis or risk prediction. Mental-health language is used here as a demanding longitudinal weak-measurement case study in which lexical, sentiment, and neural measurements overlap but are not interchangeable and are affected by changing corpus conditions. The resulting Reddit trajectory is therefore interpreted as a relative latent state of depression-related discourse measurements within the analyzed corpus, not as population prevalence or individual clinical status.

3 DynaLDI

3.1 Dynamic weak-measurement model

Let $t \in \{1, \ldots, T\}$ index time and $m \in \{1, \ldots, M\}$ index weak measurement sources. Source m produces observation $y_{m,t}$ of an unobserved scalar state d_t. We additionally observe nuisance covariates $\mathbf{c}_t$ describing the measurement process and reliability covariates $\mathbf{u}_t$ describing conditions under which measurement precision may change. DynaLDI specifies

$$y_{m,t} = \alpha_m + \boldsymbol{\gamma}_m^\top \mathbf{c}_t + \beta_m d_t + \epsilon_{m,t}, \tag{1}$$

$$d_t = \phi d_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, q), \tag{2}$$

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t), \qquad \mathbf{R}_t = \mathrm{diag}(R_{1,t}, \ldots, R_{M,t}),$$

$$\log R_{m,t} = \log r_m + \mathbf{a}_m^\top \mathbf{u}_t. \tag{3}$$

Here α_m is a source-specific intercept, $\boldsymbol{\gamma}_m$ captures nuisance dependence, β_m is the loading on the shared state, and $R_{m,t}$ is the source's time-varying conditional measurement variance. The loading vector is static: temporal changes in source influence arise through measurement precision rather than freely varying semantic loadings. This restriction reduces ambiguity between changes in the latent state, source orientation, and observation noise.

A variable may enter both $\mathbf{c}_t$ and $\mathbf{u}_t$ because mean drift and measurement reliability represent distinct properties of the observation process. For example, corpus volume may affect the expected value of a weak measurement while independently changing its precision. The corresponding parameters enter different components of the likelihood and are regularized separately.

3.2 Inference, missingness, and source influence

Model parameters are estimated jointly by alternating state inference with regularized updates of nuisance effects, source loadings, state dynamics, and variance parameters. Nuisance and reliability coefficients use ridge regularization; the loading vector is normalized to unit ℓ_2 norm and oriented so that the lexical-distress-presence loading is non-negative. Detailed initialization, update rules, regularization values, numerical bounds, and robust-observation thresholds are provided in the appendix.

Isolated discordant observation days are retained rather than deleted: DynaLDI inflates their daily measurement covariance according to a multivariate residual diagnostic. When no usable measurements are available at time t, no measurement update is performed and the state is propagated only through Eq. 2; thus missing dates do not require interpolation or pseudo-observations.

Given fitted parameters, the Kalman filter produces $p(d_t \mid \mathbf{y}_{1:t})$, while Rauch–Tung–Striebel smoothing produces $p(d_t \mid \mathbf{y}_{1:T})$ (Kalman, 1960; Rauch et al., 1965). Because measurement parameters are estimated globally from the complete dataset, the filtered trajectory is observation-one-sided conditional on the fitted measurement model rather than a fully prospective estimate. The smoothed trajectory is the primary representation for retrospective analyses.

For diagonal $\mathbf{R}_t$, conditional information from source m is proportional to $I_{m,t} = \beta_m^2 / R_{m,t}$. We report the normalized information share

$$w_{m,t} = \frac{\beta_m^2 / R_{m,t}}{\sum_{j=1}^{M} \beta_j^2 / R_{j,t}}, \tag{4}$$

which quantifies a source's contribution to conditional measurement precision. These are diagnostic information shares, not mixture weights in Eq. 1.

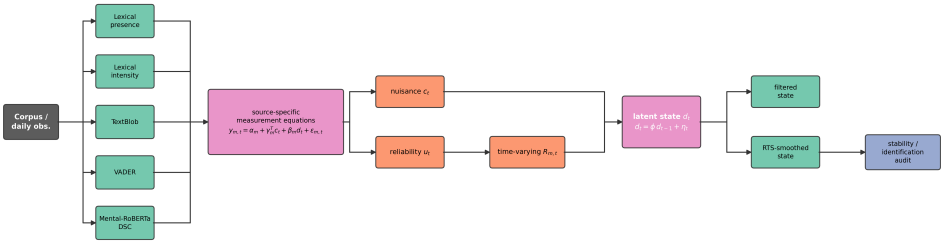


Figure 1: DynaLDI schematic diagram. Weak linguistic measurements load on a shared autoregressive latent state, with nuisance adjustment and time-varying reliability.

The latent factor is otherwise identifiable only up to sign and scale. Reported real-data trajectories are therefore standardized over observed days and oriented using the lexical-distress-presence source. Their values represent relative deviations within the analyzed corpus rather than an external clinical scale.

3.3 Identification and stability diagnostics

The Kalman/RTS posterior variance describes state uncertainty conditional on a particular fitted measurement model; it does not represent uncertainty about whether a different measurement decomposition would be obtained after re-estimation. We therefore perturb source availability, nuisance specification, temporal support, and data-quality support and distinguish full re-estimation sensitivity from fixed-parameter inference sensitivity.

Let V_t^{post} denote conditional posterior variance, V_t^{id} family-balanced squared deviation under full re-estimation, V_t^{support} the corresponding fixed-parameter support sensitivity, and V_t^{block} variability induced by random contiguous block deletion. We summarize these quantities through

$$V_t^{\text{stab}} = V_t^{\text{post}} + V_t^{\text{id}} + V_t^{\text{support}} + V_t^{\text{block}}, \tag{5}$$

$$\rho_t = \frac{V_t^{\text{post}}}{V_t^{\text{stab}}}. \tag{6}$$

Because the components are not derived as independent random variables, V_t^{stab} is not a posterior variance and its associated envelope is not presented as a Bayesian credible interval or frequentist confidence interval. Likewise, ρ_t is a *posterior-variance fraction*, not a probability of correctness or a calibrated local reliability score. Its empirical diagnostic properties are evaluated separately in Section 5.

4 Experimental Setup

4.1 Reddit case study and weak measurements

We construct a daily longitudinal panel from public `r/depression` submissions between 1 January 2009 and 31 December 2022. After deterministic parsing and record validation, the corpus contains 1,516,442 unique valid submissions from 499,067 authors. Of 5,113 calendar days, 4,897 contain at least one valid submission and 216 are retained as missing observation days. The subreddit is self-selected and is not treated as a representative population sample.

Five standardized daily weak measurements span three representation families: lexical-distress presence, lexical-distress intensity, TextBlob negativity (Loria, 2018), VADER negativity (Hutto & Gilbert, 2014), and Mental-RoBERTa Depression–Suicidality Contrast (DSC) (Ji et al., 2022). The latter is a RoBERTa-based (Liu et al., 2019) depression–suicidality contrast model trained on a separate labeled Reddit dataset of 1,895 posts (Haque et al., 2021); its held-out performance includes AUROC 0.824 and AUPRC 0.845. It is used only as a weak contrast measurement, not as a depression probability or clinical score. Additional classifier and calibration results are reported in Appendix A.2.

Daily corpus descriptors include post and unique-author counts, full-text availability, deleted and removed fractions, and observable word count. These variables enter the nuisance model, while standardized log post volume and title-only fraction parameterize time-varying measurement variance. Titles are retained for all valid records; unavailable deleted or removed bodies are not reconstructed.

4.2 KNOWN-TRUTH SIMULATIONS AND BASELINES

The primary controlled evaluation uses numerical longitudinal measurement sequences rather than generated text. Every sequence is produced from a prespecified data-generating process for which the true latent state, loadings, time-varying measurement variances, nuisance variables, and missingness are known. Each sequence has 5,113 time points, matching the Reddit calendar.

The confirmatory benchmark contains 480 held-out sequences (48 per regime) across ten conditions: clean stationary data, volume-dependent heteroskedasticity, nuisance drift, correlated source noise, loading drift, missing blocks, outliers, structural shocks, nuisance–latent correlation, and combined stress. The final DynaLDI configuration was fixed before this benchmark and no hyperparameters were selected from its results. We compare against Equal-Z, PCA/static, a raw constant-variance dynamic estimator, a nuisance-adjusted constant-variance model, earlier dynamic variants, and an oracle using the known data-generating parameters.

Primary simulation evaluation uses latent-state RMSE after resolving sign and scale ambiguity, supplemented by Pearson and Spearman correlation, paired RMSE differences, interval coverage, loading cosine similarity, and recovery of time-varying precision. A separate, previously unseen 384-sequence benchmark spanning 16 scenarios evaluates the stability diagnostics; its additional scenarios target identification failures such as weak-factor geometry, competing latent structure, reliability–state confounding, loading rotation, removal-process shift, and dominant-source variance collapse.

The pre-specified final configuration uses a diagonal conditional measurement covariance and robust per-day covariance inflation. Implementation settings and regularization are reported in Appendix A.1, while the known-truth generator, simulation regimes, evaluation design, and confirmatory acceptance criteria are described in Appendix B.

4.3 EXTERNAL TEMPORAL COMPARISONS

We evaluate whether the fixed Reddit trajectory exhibits ecological temporal alignment with four independent data families: CDC Household Pulse Survey depressive-symptom estimates, NSDUH adult major depressive episode estimates, Google Trends relative search volume for the literal English query `depression` in the Health category (Worldwide and United States), and U.S. age-adjusted suicide mortality from NVSS (National Center for Health Statistics, 2026; Substance Abuse and Mental Health Services Administration, 2020; Google, 2026; Garnett & Curtin, 2024). None is treated as ground truth for the Reddit latent state. The primary comparisons cover 52 CDC survey periods during 2020–2022, 10 annual NSDUH observations during 2010–2019, 156 monthly observations for each Google Trends series during 2010–2022, and 13 annual NVSS observations during 2010–2022. For each target interval we aggregate the fixed daily DynaLDI and Equal-Z trajectories over corresponding observed Reddit days. Target definitions, directions, coverage rules, and lag zero as the primary alignment were specified before association analysis; DynaLDI is never refit or retuned to an external series. We report Pearson and Spearman correlation, standardized RMSE (zRMSE), directional agreement, trend-robustness analyses, and paired block-bootstrap uncertainty. U.S.-specific targets additionally have a geographic mismatch because the Reddit corpus is not geographically restricted. These analyses test ecological temporal convergence, not clinical or individual-level validity.

4.4 COMPUTATIONAL BENCHMARK

We evaluate the state-space computational kernel on all 30 combinations of

$$T \in \{500, 1000, 2500, 5000, 10000, 20000\}, \qquad M \in \{3, 5, 10, 20, 50\}.$$

The benchmark records sufficient-statistic construction, filtering, RTS smoothing, peak resident memory, deterministic reruns, and float32-versus-float64 agreement. A prespecified cross-section

additionally repeats likelihood optimization for (ϕ, q) with data-generating measurement parameters held fixed. This benchmark characterizes the state-space kernel rather than the complete joint DynaLDI optimizer; absolute timing is therefore hardware dependent.

5 RESULTS

5.1 KNOWN-TRUTH LATENT-STATE RECOVERY

Table 1 summarizes state recovery over the pre-specified confirmatory benchmark of 480 held-out known-truth simulated longitudinal sequences. DynaLDI achieves mean RMSE 0.460, Pearson correlation 0.881, and Spearman correlation 0.874. Equal-weight standardized aggregation (Equal-Z) obtains RMSE 0.606, corresponding to a 24.0% reduction in latent-state RMSE for DynaLDI. Relative to the earlier dynamic estimator (RMSE 0.495), DynaLDI reduces RMSE by 7.0% and has lower RMSE in 91.9% of paired trials; the mean paired improvement is 0.0349 with 95% confidence interval $[0.0266, 0.0431]$.

The improvements are not confined to a single matched condition. DynaLDI reaches RMSE 0.362 under nuisance drift, compared with 0.601 for Equal-Z and 0.626 for the earlier dynamic estimator. Under volume-dependent heteroskedasticity and block missingness it reaches RMSE 0.369 and 0.371, respectively, close to the corresponding oracle values of 0.361 and 0.363. These results support the intended roles of joint nuisance adjustment, reliability-conditioned variance, and temporal state propagation. The stress regimes also define clear limits. Under correlated source noise, DynaLDI obtains RMSE 0.597 versus 0.677 for Equal-Z, but remains well above the oracle value 0.472. Under combined stress, DynaLDI reaches 0.872 compared with 0.861 for Equal-Z; this is the only regime in which Equal-Z has lower mean RMSE. Thus the diagonal conditional covariance model is not robust to arbitrary residual source dependence or simultaneous severe misspecification. DynaLDI also recovers the simulated measurement structure accurately. Mean cosine similarity between estimated and true loading vectors is 0.9991, and correlation between estimated and true time-varying measurement precision is 0.9847. Nominal 95% state intervals achieve 94.9% empirical coverage across matched and near-matched regimes. Coverage deteriorates to 0.654 under correlated source noise and 0.489 under combined stress, showing that posterior uncertainty is calibrated conditional on an approximately adequate measurement model rather than against arbitrary model misspecification. Full regime-level results, coverage diagnostics, and filtered-versus-smoothed comparisons are provided in the appendix.

5.2 LONGITUDINAL REDDIT APPLICATION

Fitting DynaLDI to the five-source daily Reddit panel yields estimated persistence $\hat{\phi} = 0.956$ and innovation variance $\hat{q} = 0.00165$. The primary representation is the retrospective RTS-smoothed state, standardized over observed days and oriented toward the lexical-distress-presence direction. Figure 2 shows the resulting trajectory and conditional posterior uncertainty.

The fitted loading vector is approximately $\hat{\boldsymbol{\beta}} = [0.539,\ 0.122,\ -0.124,\ 0.133,\ 0.814]^\top$ for lexical presence, lexical intensity, TextBlob negativity, VADER negativity, and Mental-RoBERTa DSC, respectively. Because effective source influence depends on both loading and conditional precision, the median information shares are 42.9% for Mental-RoBERTa DSC, 35.8% for lexical presence, 18.7% for lexical intensity, 0.8% for TextBlob, and 0.7% for VADER. The median effective source count is approximately 2.8, and no observed day assigns more than 80% of total conditional information to one source. The inferred trajectory is therefore not simply a rescaled transformer score. Source-level measurement precision, fitted loadings, and median information-share summaries are reported in Appendix A.3. Relative to Equal-Z, DynaLDI is also less strongly associated with several observable corpus characteristics over the denser 2012–2022 reporting period, excluding the markedly sparse early sampling regime. Pearson correlation with full-text fraction decreases from 0.810 to 0.225, with mean observable word count from 0.773 to 0.214, and with log post count from -0.647 to -0.134. Correlation with removed fraction remains substantial at -0.456, however. These reductions therefore indicate weaker mechanical association with several observed properties of the measurement process; they do not establish complete deconfounding or imply that the latent state is independent of corpus composition.

Table 1: Known-truth state-recovery performance over 480 held-out simulated longitudinal sequences. Lower RMSE is better; higher correlations are better. Bold marks the best non-oracle result.

Method	RMSE ↓	Pearson ↑	Spearman ↑
Oracle	0.400	0.918	0.911
DynaLDI	**0.460**	**0.881**	**0.874**
Covariance-aware dynamic	0.474	0.871	0.866
Earlier dynamic estimator	0.495	0.863	0.855
Nuisance constant-variance	0.518	0.850	0.846
Equal-Z	0.606	0.811	0.809
Raw dynamic	0.643	0.733	0.726
PCA/static	0.748	0.673	0.670

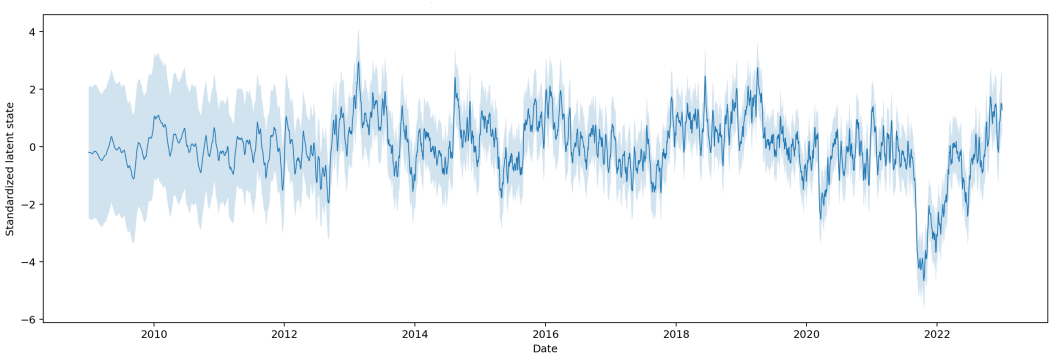


Figure 2: Estimated 2009–2022 DynaLDI trajectory for `r/depression`. The curve represents the corpus-standardized relative latent state jointly supported by the five depression-related linguistic measurements; shading denotes the conditional 95% state interval. Values are not estimates of clinical prevalence, diagnosis, or individual risk.

5.3 Conditional state stability and measurement-model identification

The real-data perturbation audit reveals a clear distinction between state inference conditional on the fitted measurement mapping and re-identification of that mapping. Fixed-model temporal restrictions remain highly correlated with the reference trajectory, whereas full re-estimation on the same supports can produce substantially different trajectories. Sensitivity is also greater when high-removal-content days are excluded. Exact refitting and small parameter perturbations reproduce the reference solution, arguing against ordinary numerical instability. Full perturbation results are reported in Appendix C.1.

The composite diagnostic likewise indicates that conditional posterior uncertainty captures only part of the observed sensitivity. Median posterior-variance fraction is $\rho_t = 0.479$; mean variance shares are approximately 46.0% posterior, 36.1% identification, 13.9% fixed-support, and 4.1% block-deletion sensitivity. The full decomposition is reported in Appendix C. Because the true Reddit state is unknown, we evaluate the diagnostics on 384 held-out simulated sequences comprising 24 replicates of each of 16 known-truth scenarios, including matched conditions and targeted identification failures. Table 2 reports the pre-specified diagnostic-performance and benchmark-integrity checks.

The diagnostics are weak pointwise error indicators: $1-\rho_t$ has median Spearman correlation 0.0358 with absolute state error, and identification RMS is similarly weak locally. Their behavior is more informative after aggregation, with scenario-level identification-RMS correlation of 0.759 with true RMSE. The full sensitivity envelope improves severe-misspecification coverage from 0.7157 to 0.8435, but is conservative under matched conditions, where coverage increases from 0.9347 to 0.9811. We therefore interpret identification sensitivity as a dataset- or regime-level model-health diagnostic and the full envelope as a sensitivity-aware expansion rather than a replacement for the conditional posterior. Detailed validation criteria and scenario-level results are reported in Appendix D.

Table 2: Central stability-validation results. Four of eight pre-specified diagnostic-performance criteria pass, while all four benchmark-integrity checks pass. The latter verify benchmark behavior rather than diagnostic accuracy.

Quantity	Result
Diagnostic-performance criteria passed	4 / 8
Benchmark-integrity checks passed	4 / 4
Spearman: $1 - \rho_t$ vs. $\lvert e_t \rvert$	0.0358
Matched posterior / full-envelope coverage	0.9347 / 0.9811
Severe posterior / full-envelope coverage	0.7157 / 0.8435
Scenario-level Spearman: identification RMS vs. RMSE	0.759

5.4 External temporal convergence

Figure 3 summarizes the prespecified lag-zero comparisons. Across all five external series, DynaLDI has higher Pearson and Spearman correlations and lower zRMSE than Equal-Z, although the strength and interpretation of convergence differ by construct.

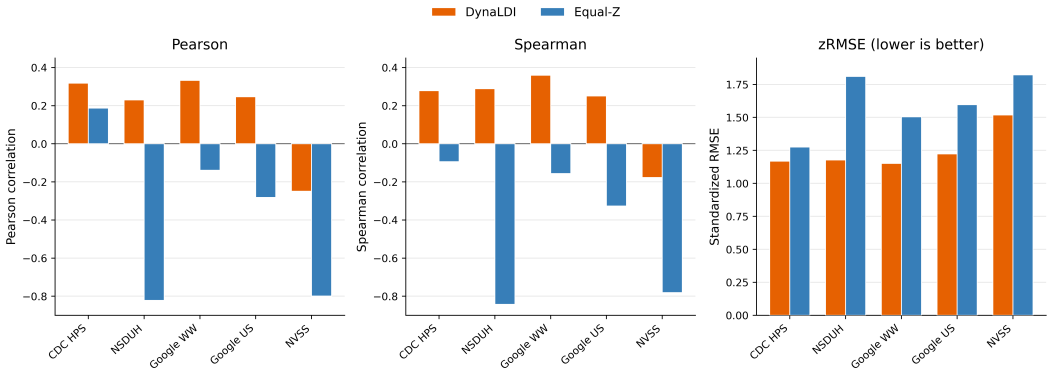


Figure 3: Prespecified lag-zero ecological comparisons of DynaLDI and Equal-Z across five external series. Bars show Pearson correlation, Spearman correlation, and zRMSE; uncertainty is assessed separately by paired block bootstrap.

CDC HPS shows modest positive alignment, but bootstrap correlation intervals include zero, so the evidence is suggestive. NSDUH is positive at lag zero (Pearson 0.230, Spearman 0.289), but the association largely disappears after detrending, indicating a predominantly long-horizon trend relationship. Google Trends provides the clearest trend-robust convergence: for the Worldwide series, DynaLDI obtains Pearson 0.333 and Spearman 0.360 and remains positively associated after first differencing and detrending; the U.S. series shows the same qualitative pattern. In contrast, U.S. age-adjusted suicide mortality from NVSS remains negatively associated with DynaLDI (Pearson -0.249, Spearman -0.177), providing a construct boundary rather than positive validation. We therefore interpret the results as selective ecological temporal convergence rather than evidence of clinical validity. Full bootstrap intervals, target-specific metrics, transformations, and sensitivity analyses are reported in Appendix E.

5.5 Computational scaling and reproducibility

Across 30 prespecified (T, M) configurations, state-space kernel runtime scales approximately linearly with sequence length, with fitted exponent 0.9995 ($R^2 = 0.9997$). At the largest tested configuration ($T = 20{,}000$, $M = 50$), runtime is 0.082 seconds and peak resident memory is approximately 238 MiB. All deterministic reruns reproduce the smoothed state exactly, and numerical-precision checks pass. These results characterize the scalar state-space kernel rather than the complete joint optimizer; the full scaling grid, component timings, precision checks, and (ϕ, q) likelihood experiment are reported in Appendix F.

6 Discussion and Limitations

DynaLDI treats longitudinal weak-signal aggregation as a measurement problem rather than fixed averaging. Known-truth experiments show that jointly modeling temporal persistence, nuisance variation, source loadings, and time-varying measurement precision improves latent-state recovery over Equal-Z and simplified dynamic alternatives. The improvement under nuisance drift also indicates that the gain is not attributable to temporal smoothing alone. A central result is the distinction between conditional state uncertainty and measurement-model identification. State inference can remain stable once the measurement mapping is fixed even when re-estimation under altered sources or temporal support produces substantially different decompositions. The Kalman/RTS posterior does not represent this latter sensitivity because it conditions on the fitted model. Perturbation diagnostics expose it, but their known-truth evaluation also limits their interpretation: $1 - \rho_t$ and identification RMS are weak predictors of instantaneous state error and are more informative after aggregation to datasets or regimes. The composite sensitivity envelope improves coverage under severe misspecification but is conservative under matched conditions; it therefore complements rather than replaces conditional posterior uncertainty. Several assumptions constrain the method. The conditional measurement covariance is diagonal: correlated source noise degrades recovery and coverage, and combined stress is the only state-recovery regime in which Equal-Z has lower mean RMSE. Static loadings improve identifiability while allowing effective source influence to vary through precision, but cannot represent unrestricted semantic drift. The Reddit application is self-selected, platform-specific, and affected by participation, moderation, deletion, and changing text availability. Removed content is unobserved and is an important source of sensitivity. The five inputs are weak linguistic measurements rather than clinical observations, and Mental-RoBERTa DSC is a depression–suicidality contrast rather than a depression probability. Because the DSC training corpus and the longitudinal corpus both contain `r/depression` posts, some post-level overlap cannot be ruled out; DSC is therefore treated only as one weak measurement rather than as independent ground truth. The late-2021 trough changes substantially under measurement-model re-identification and removal-related support restrictions, so we assign it no historical or clinical interpretation. External comparisons provide selective ecological convergence rather than ground truth. CDC alignment is suggestive, NSDUH alignment is largely trend-driven, Google Trends is more robust to detrending and differencing, and suicide mortality shows no positive contemporaneous convergence. U.S.-specific targets also have a geographic mismatch with the unrestricted Reddit corpus. Finally, the computational benchmark characterizes the state-space kernel rather than the full joint optimizer; numerical convergence and reproducibility should not be conflated with statistical identification. The interpretation boundaries applying to these results are summarized formally in Appendix G.

7 Conclusion

We introduced DynaLDI, a dynamic latent-index model for recovering a shared longitudinal state from heterogeneous weak linguistic measurements under nuisance variation, missing observations, and time-varying reliability. Known-truth simulations show improved state recovery over fixed aggregation and simplified dynamic alternatives, while the Reddit application demonstrates why conditional state uncertainty must be distinguished from measurement-model identification sensitivity. A separate known-truth benchmark shows that perturbation-derived diagnostics are more useful for characterizing regime-level instability and sensitivity-aware coverage than for predicting instantaneous error. Prespecified external comparisons provide selective ecological temporal convergence, with the strongest trend-robust evidence from Google Trends and a negative contemporaneous result for suicide mortality. DynaLDI is therefore intended as a framework for constructing and auditing longitudinal weak-signal indices while making their measurement assumptions and identification limits explicit. The Reddit trajectory remains a relative latent state of depression-related discourse measurements, not an estimate of clinical prevalence, diagnosis, population mental-health burden, or individual risk.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. We analyze public `r/depression` submissions (2009–2022) from the Pushshift Reddit dataset (Baumgartner et al., 2020), obtained via The Eye's `redarcs` archive (The Eye, n.d.). Because the longitudinal Reddit corpus contains mental-health disclosures, including references to suicidality, all DynaLDI longitudinal analyses and reported Reddit outputs are restricted to daily aggregate measurements. We do not build user profiles, make individual-level predictions, or attempt to identify, contact, or track any user, and the paper reproduces no post text, usernames, or post identifiers. Deleted or removed post bodies are not reconstructed. Mental-RoBERTa DSC is trained on the 1,895-post SDCNL labeled dataset (Haque et al., 2021), labeled by subreddit membership rather than clinical assessment, and serves only as one weak measurement. The external comparators (CDC Household Pulse Survey, NSDUH, Google Trends, and NVSS) are published population-level statistics without individual records.

The main risks are misinterpretation and misuse. The inferred trajectory is a relative state of depression-related discourse within a self-selected, platform-specific community, measured partly with English-language lexicon and sentiment tools. It is not a measure of depression prevalence, diagnosis, or suicide risk, and its negative association with U.S. suicide mortality shows why it should not be treated as one. It should not be used to screen, monitor, or make decisions about individuals or specific communities. Appendix G lists the interpretation boundaries that apply to all results.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A.1 describe the model, identification conventions, initialization and fitting procedure, numerical bounds, measurement-model updates, missing-data handling, and final estimator configuration (Table 3). Corpus construction, the complete lexical vocabulary, weak-measurement definitions, and the training task, archived training configuration, held-out evaluation, and calibration of Mental-RoBERTa DSC are described in Appendix A.2. The known-truth generator structure, principal parameters, deterministic seeding rule, replicate identifiers, baselines, evaluation metrics, and pre-specified state-recovery checks are reported in Appendix B. The independent stability-validation benchmark is described in Appendix D, and the real-data perturbation audit in Appendix C. Exact refitting and parameter-perturbation checks are reported in Appendix C.1. External target definitions, primary periods, alignment rules, metric definitions, bootstrap designs, and target-specific robustness analyses are provided in Appendix E, with preprocessing integrity checks in Appendix E.1. The computational benchmark, including deterministic reruns and numerical-precision checks, is reported in Appendix F. Interpretation boundaries applying to all analyses are summarized in Appendix G.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, including drafting and editing the manuscript. We also used them to provide feedback on research methodology and experimental design, to implement and debug parts of the analysis code, to check our interpretation of results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pp. 830–839, 2020. doi: 10.1609/icwsm.v14i1.7347.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 152–162. Association for Computational Linguistics, 2017. URL `https://aclanthology.org/E17-1015/`.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006. doi: 10.1145/1143844.1143859.

Sy-Miin Chow, Jiyun Zu, Kim Shifren, and Guangjian Zhang. Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research*, 46(2):303–339, 2011. doi: 10.1080/00273171.2011.563697.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 51–60, 2014. doi: 10.3115/v1/W14-3207.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pp. 31–39, 2015. doi: 10.3115/v1/W15-1204.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pp. 128–137, 2013. doi: 10.1609/icwsm.v7i1.14432.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6(12):e26752, 2011.

James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2 edition, 2012. doi: 10.1093/acprof:oso/9780199641178.001.0001.

Wayne A. Fuller. *Measurement Error Models*. Wiley, 1987.

Matthew F. Garnett and Sally C. Curtin. Suicide mortality in the united states, 2002–2022. NCHS Data Brief, No. 509, 2024. URL `https://www.cdc.gov/nchs/products/databriefs/db509.htm`.

Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

Google. Google trends, 2026. URL `https://trends.google.com/trends/`. Relative search-volume data.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1489–1501, 2016. doi: 10.18653/v1/P16-1141.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, volume 12895 of *Lecture Notes in Computer Science*, pp. 436–447. Springer, 2021. doi: 10.1007/978-3-030-86383-8_35.

Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*, 2022.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, pp. 5574–5584, 2017.

Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 675–684, 2015. doi: 10.1145/2783258.2783277.

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

Steven Loria. Textblob documentation. `https://textblob.readthedocs.io/`, 2018.

National Center for Health Statistics. Household pulse survey: Anxiety and depression. Centers for Disease Control and Prevention, 2026. URL `https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm`. National estimates of anxiety and depressive-disorder symptoms.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pilar Poncela, Esther Ruiz, and Karen Miranda. Factor extraction using kalman filter and smoothing: This is not just another survey. *International Journal of Forecasting*, 37(4):1399–1425, 2021. doi: 10.1016/j.ijforecast.2021.01.027.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Alexander J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Herbert E. Rauch, F. Tung, and Charlotte T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*, 7:13006, 2017. doi: 10.1038/s41598-017-12961-9.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

Substance Abuse and Mental Health Services Administration. Results from the 2019 national survey on drug use and health: Detailed tables. Center for Behavioral Health Statistics and Quality, 2020. URL `https://www.samhsa.gov/data/report/2019-nsduh-detailed-tables`.

The Eye. redarcs: Reddit archives, 2005–2022, n.d. URL `https://the-eye.eu/redarcs/`. Accessed January 18, 2024.

A Additional Implementation and Data Details

This appendix provides implementation, data-construction, experimental, and robustness details omitted from the main paper for space. All confirmatory analyses use the pre-specified DynaLDI configuration. External targets are not used to fit or retune the Reddit model.

A.1 Estimator implementation

DynaLDI uses the measurement and state equations defined in Section 3, with a scalar latent state, static unit-norm loading vector $\boldsymbol{\beta}$, source-specific nuisance effects, and reliability-conditioned diagonal observation covariance $\mathbf{R}_t$.

The factor scale and orientation are identified by

$$\|\boldsymbol{\beta}\|_2 = 1, \tag{7}$$

with the sign chosen so that the lexical-distress-presence (De Choudhury et al., 2013; Coppersmith et al., 2014), loading is non-negative. Source influence varies through conditional precision,

$$I_{m,t} = \frac{\beta_m^2}{R_{m,t}}, \qquad w_{m,t} = \frac{I_{m,t}}{\sum_j I_{j,t}}, \tag{8}$$

rather than through unconstrained time-varying loadings.

Initialization first fits source-specific ridge nuisance regressions. The first singular-vector factor of the nuisance-adjusted panel initializes the loading direction and latent state. Initial ϕ is obtained from adjacent observed state estimates and q from the corresponding innovation variance.

The initial persistence estimate is clipped to $[0.05, 0.98]$, and the initial innovation variance is clipped to $[10^{-4}, 5]$.

Measurement coefficients are updated by expected weighted least squares using the smoothed latent-state moments

$$\mathbb{E}[d_t] = m_t, \qquad \mathbb{E}[d_t^2] = m_t^2 + P_t. \tag{9}$$

Observation weights are proportional to the inverse conditional measurement variance and are clipped to $[10^{-6}, 10^6]$. Nuisance coefficients receive the ridge penalty reported in Table 3, while the source intercept and latent-state loading are left unpenalized. After each measurement-model update, $\boldsymbol{\beta}$ is renormalized to unit ℓ_2 norm and the latent state mean and variance are rescaled consistently.

The final fitting procedure alternates nuisance and loading updates, reliability-conditioned variance estimation, robust day-level covariance inflation, Kalman filtering, RTS smoothing, and optimization of (ϕ, q). Eight joint iterations are performed before the final measurement and state update.

The final numerical configuration is summarized in Table 3.

For source residual $e_{m,t}$, reliability estimation uses the robust scale

$$s_m = 1.4826 \text{ median } |e_{m,t} - \text{median}(e_{m,t})|, \tag{10}$$

and models log-squared residual magnitude as a function of standardized log post volume and title-only fraction. Reliability slopes are ridge-regularized, bounded as shown in Table 3, and damped between successive joint iterations. Base measurement variance is estimated after limiting only the extreme squared-residual tail at the 99.5th percentile when sufficient observations are available.

Discordant days are retained rather than deleted. For the nuisance-adjusted residual vector $\mathbf{e}_t$,

$$M_t = \mathbf{e}_t^\top \mathbf{R}_t^{-1} \mathbf{e}_t, \tag{11}$$

and

$$\kappa_t = \text{clip}\left(\frac{M_t}{18}, 1, 10\right). \tag{12}$$

The observation covariance is then replaced by $\kappa_t \mathbf{R}_t$, reducing the conditional precision of unusually discordant observations without clipping their measured values.

Table 3: Final DynaLDI implementation settings.

Setting	Value
Joint fitting iterations	8
Nuisance ridge penalty	150
Variance-slope ridge penalty	400
Variance-slope damping	0.45
Reliability-slope bounds	$[-1.75, 1.75]$
Log variance-adjustment bounds	$[-6, 6]$
Base-variance bounds	$[10^{-5}, 100]$
Day-specific variance bounds	$[10^{-7}, 10^{5}]$
Robust inflation threshold	18
Maximum covariance inflation	$10\times$
State-persistence bound	$\lvert\phi\rvert \leq 0.995$
$\log q$ optimization bounds	$[-12, 4]$
Optimizer	L-BFGS-B
Optimizer maximum iterations	180
Loading normalization	Unit ℓ_2 norm
Primary state output	RTS-smoothed
Missing-day handling	Prediction only
Primary random seed	42

Because the latent state is scalar and $\mathbf{R}_t$ is diagonal, the multi-source measurement system can be reduced at each time point to

$$J_t = \boldsymbol{\beta}^\top \mathbf{R}_t^{-1} \boldsymbol{\beta}, \qquad V_t = J_t^{-1}, \tag{13}$$

and

$$z_t = V_t \boldsymbol{\beta}^\top \mathbf{R}_t^{-1} \widetilde{\mathbf{y}}_t, \tag{14}$$

where $\widetilde{\mathbf{y}}_t$ is the nuisance-adjusted measurement vector. Filtering and smoothing then operate on these scalar sufficient statistics.

Completely missing dates receive no pseudo-observation or interpolated measurement. Their state is propagated only through

$$m_{t|t-1} = \phi m_{t-1|t-1}, \tag{15}$$

$$P_{t|t-1} = \phi^2 P_{t-1|t-1} + q. \tag{16}$$

The final Reddit fit gives

$$\hat{\phi} = 0.955532, \qquad \hat{q} = 0.001653. \tag{17}$$

A.2 Corpus and weak linguistic measurements

The canonical Reddit corpus spans 1 January 2009 through 31 December 2022. Its principal characteristics are shown in Table 4.

Deleted and removed body placeholders are excluded from linguistic text; their availability status is retained as metadata. The nuisance panel contains standardized log post count, log unique-author count, full-text fraction, deleted fraction, removed fraction, and mean observable word count.

The five final weak measurements are lexical-distress presence, lexical-distress intensity, TextBlob negativity, VADER negativity, and Mental-RoBERTa DSC.

The fixed lexical vocabulary is: `depression`, `depressed`, `depressive`, `hopeless`, `hopelessness`, `worthless`, `worthlessness`, `anhedonia`, `anhedonic`, `suicide`, `suicidal`, `self harm`, `self-harm`, `want to die`, `kill myself`, and `no reason to live`. Matching is case-insensitive. Lexical presence is the daily fraction of posts containing at least one match, whereas lexical intensity is 100 times the total number of matches divided by total observable words.

TextBlob and VADER are sign-reversed so that larger standardized values represent greater negativity.

Table 4: Canonical Reddit corpus used to construct the daily panel.

Quantity	Value
Valid unique submissions	1,516,442
Unique authors	499,067
Calendar days	5,113
Observed days	4,897
Completely missing days	216
Full-text records	929,210
Title-only records	587,232
Deleted or removed body fraction	36.66%

Mental-RoBERTa DSC is trained on 1,895 labeled examples from the `r/depression` versus `r/SuicideWatch` contrast. The held-out test set contains 379 examples. Test accuracy is 0.7520, macro-F1 is 0.7501, AUROC is 0.8239, and AUPRC is 0.8454. Post-hoc temperature scaling uses the frozen fitted value $T = 1.818269$ and improves Brier score from 0.229 to 0.200 and negative log-likelihood from 0.999 to 0.633. The archived final classifier records the training configuration summarized in Table 5. DSC is used only as a weak depression–suicidality contrast measurement and is not interpreted as a probability of depression, suicidality, diagnosis, symptom severity, or individual clinical status.

Table 5: Mental-RoBERTa DSC training and calibration configuration recovered from the archived final model.

Setting	Value
Architecture	RoBERTa sequence classifier
Transformer layers	12
Hidden size	768
Attention heads	12
Maximum token length	512
Training epochs	6
Train batch size	8
Evaluation batch size	16
Gradient accumulation	1
Learning rate	2×10^{-5}
Weight decay	0.01
Optimizer	AdamW (fused)
Adam β_1, β_2	0.9, 0.999
Adam ϵ	10^{-8}
Maximum gradient norm	1.0
Learning-rate schedule	Linear
Warmup ratio	0.10
Training precision	FP16
Random seed	42
Evaluation frequency	Every epoch
Checkpoint frequency	Every epoch
Checkpoint selection	Best macro-F1
Temperature scaling	$T = 1.818269$

All five daily measurements are standardized over observed days before DynaLDI fitting. Corpus-association summaries use 2012–2022 because the earliest corpus years have markedly different sampling density: mean daily post counts increase from 1.70 in 2009, 6.21 in 2010, and 22.21 in 2011 to 74.88 in 2012 and substantially higher levels thereafter. This restriction applies only to the descriptive corpus-association diagnostic; the DynaLDI model and reported latent trajectory use the full 2009–2022 corpus.

A.3 Final Reddit measurement model

Table 6 reports the fitted source loadings and median conditional information shares.

Table 6: Final Reddit measurement-model summary. Median information shares are computed independently across days and therefore need not sum exactly to one.

Source	Loading	Median info. share
Lexical presence	.5388	.358
Lexical intensity	.1219	.187
TextBlob negativity	-.1241	.008
VADER negativity	.1327	.007
Mental-RoBERTa DSC	.8135	.429

The median effective number of contributing sources is approximately 2.80. The large DSC loading therefore does not imply that DSC alone determines the state; contribution depends jointly on loading magnitude and day-specific conditional precision.

Associations with observable corpus variables are reduced relative to Equal-Z but are not eliminated. Over 2012–2022, correlation with log post count changes from -0.647 to -0.134, with full-text fraction from 0.810 to 0.225, with deleted fraction from -0.588 to 0.042, and with mean observable word count from 0.773 to 0.214. The removed-fraction association remains substantial at approximately -0.456, motivating the removal-related sensitivity analysis below.

B Known-Truth State-Recovery Benchmark

This section specifies the known-truth state-recovery benchmark used to evaluate the pre-specified DynaLDI estimator. It describes the simulation design, stress regimes, comparison methods, and confirmatory criteria, and then reports regime-level recovery and interval-coverage behavior.

B.1 Simulation specification

The primary confirmatory benchmark contains

$$10 \text{ regimes} \times 48 \text{ replicates} = 480$$

previously unseen numerical longitudinal sequences. Each contains $T = 5{,}113$ time points and $M = 5$ measurement channels, matching the real-data temporal dimension. These are simulated numerical measurement sequences, not generated Reddit text.

The confirmatory benchmark uses replicate identifiers $r \in \{48, \ldots, 95\}$. With zero-based scenario index $s \in \{0, \ldots, 9\}$, the deterministic random seed is

$$42 + 100000s + r. \tag{18}$$

The base latent process is

$$d_t = 0.86 d_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(0, 1 - 0.86^2\right). \tag{19}$$

The base loading vector is

$$\boldsymbol{\beta}_0 = [0.62,\ 0.40,\ 0.52,\ 0.58,\ 0.60]^\top, \tag{20}$$

and the base measurement variances are

$$\mathbf{r} = [0.32,\ 0.58,\ 0.44,\ 0.40,\ 0.38]^\top. \tag{21}$$

The generator also produces corpus-like nuisance variables representing post volume, author count, text availability, deleted and removed fractions, and observable text length. Depending on the regime, these variables alter measurement means, conditional variances, or both.

The ten regimes are summarized in Table 7.

The correlated-noise regime uses substantial contemporaneous residual correlation, including stronger TextBlob–VADER and lexical-presence–DSC dependence. Loading drift gradually weakens the DSC and lexical-presence loadings while modestly increasing the VADER loading. The

Table 7: Known-truth simulation regimes.

Regime	Primary stress
Clean	Matched stationary measurement model
Heteroskedastic	Reliability-dependent measurement variance
Nuisance drift	Corpus-process effects on measurement means
Correlated noise	Off-diagonal residual dependence
Loading drift	Time-varying source loading strength
Missing blocks	Random and contiguous missing dates
Outliers	Sparse large source-level residuals
Structural shocks	Abrupt latent innovations
Nuisance–latent corr.	Nuisance variables correlated with state
Combined stress	Multiple misspecifications simultaneously

missingness regime combines approximately 3.5% isolated missing dates with three contiguous missing blocks. The outlier regime adds sparse source-level residuals of approximately four to eight conditional standard deviations. Structural-shock sequences receive four large latent innovations at separated points in the time series. Combined stress applies these mechanisms jointly.

The benchmark compares DynaLDI with Equal-Z, PCA/static aggregation, a raw constant-variance dynamic model, a nuisance-adjusted constant-variance model, the earlier dynamic estimator, a covariance-aware development variant, and an oracle using known data-generating measurement parameters.

The primary metric is latent-state RMSE after resolving sign and scale ambiguity. Secondary metrics are Pearson and Spearman correlation, paired RMSE differences, nominal-interval coverage, loading-vector cosine similarity, and recovery of time-varying measurement precision.

B.2 Regime-level results and confirmatory checks

Overall method-level performance is reported in Table 1 of the main paper. Figure 4 shows the corresponding regime-level RMSE and interval-coverage behavior.

DynaLDI achieves particularly strong recovery under nuisance drift, heteroskedasticity, and missing blocks. Correlated source noise produces larger error and undercoverage, while combined stress is the only regime in which Equal-Z has slightly lower mean RMSE. Mean nominal 95% coverage is 94.9% over the matched and near-matched regimes, but falls to approximately 65.4% under correlated source noise and 48.9% under combined stress.

The pre-specified confirmatory checks are summarized in Table 8. All 12 pass.

Table 8: Pre-specified acceptance checks for the 480-sequence benchmark.

Criterion	Observed	Requirement
RMSE difference vs. earlier dynamic estimator	-0.0349	< 0
RMSE difference vs. Equal-Z	-0.1454	< 0
Pearson advantage vs. earlier estimator	$+0.0188$	> 0
Paired RMSE win fraction	.9188	$\geq .75$
Lower CI bound of paired RMSE improvement	.0266	> 0
ϕ boundary-hit fraction	.000	$\leq .05$
Mean loading cosine similarity	.9991	$\geq .95$
Worst-regime loading cosine similarity	.9945	$\geq .90$
Mean precision-recovery correlation	.9847	$\geq .90$
Matched/near-matched 95% coverage	.9486	$[.90, .99]$
Worst-regime RMSE excess over Equal-Z	.0107	$\leq .05$
Structural-shock recovery noninferiority	.0156	$\geq -.02$

These criteria evaluate the already specified estimator; no hyperparameters are selected from the 480 confirmatory sequences.

Across the 480-sequence benchmark, the observation-one-sided filtered trajectory has overall RMSE 0.499 and Pearson correlation 0.863, compared with RMSE 0.460 and Pearson correlation 0.881 for the RTS-smoothed trajectory. Retrospective smoothing therefore reduces aggregate RMSE by approximately 7.8%. The filtered trajectory uses observations only through time t, but remains conditional on measurement parameters estimated from the complete sequence and is therefore not interpreted as a fully prospective estimate.

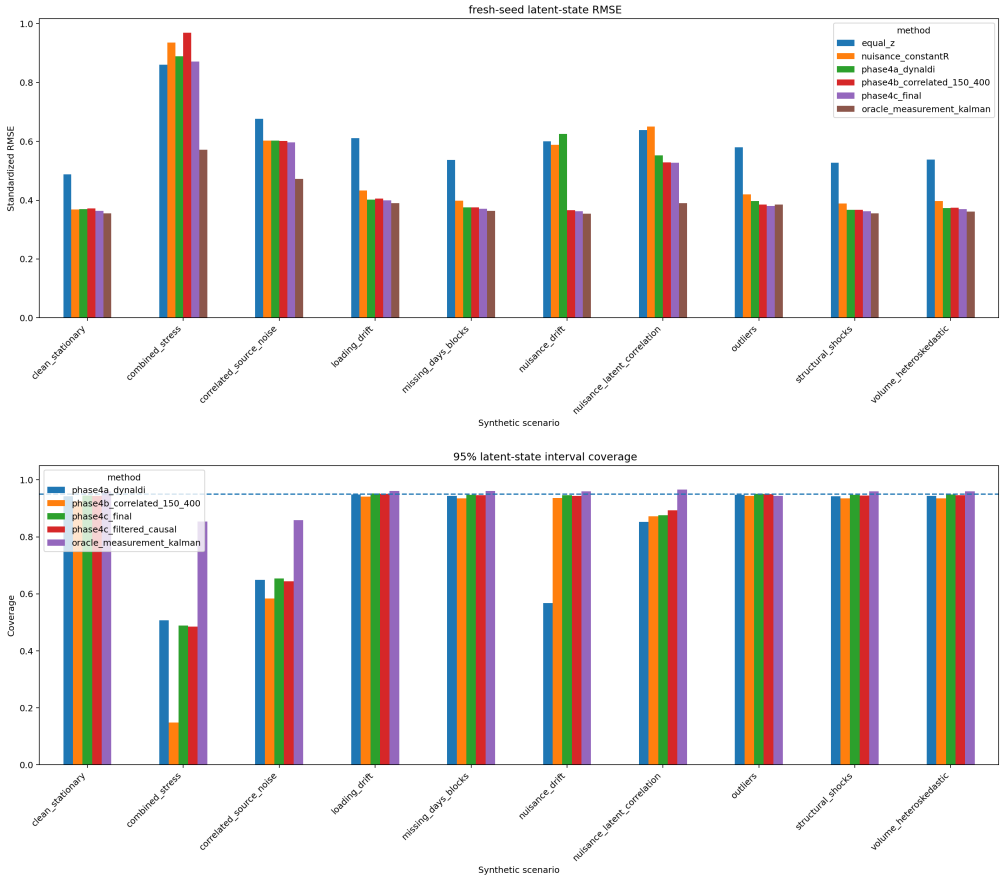


Figure 4: Known-truth simulation results by regime. Top: latent-state RMSE. Bottom: empirical coverage of nominal 95% conditional state intervals.

C Real-Data Identification and Sensitivity

This section examines the sensitivity of the fitted Reddit trajectory to changes in measurement sources, nuisance specification, temporal support, and data-quality support. It distinguishes fixed-model state sensitivity from full measurement-model re-identification and summarizes both the composite diagnostic and the late-2021 sensitivity case.

C.1 Perturbation audit

The real-data audit separates conditional state stability from measurement-model re-identification. Perturbations include leave-one-source removal, nuisance-variable removal, data-quality restrictions, temporal truncation, small parameter perturbations, and random contiguous block deletion.

Representative correlations with the original DynaLDI trajectory are reported in Table 9.

Conditional state inference is therefore substantially more stable than complete re-estimation of the measurement decomposition.

Exact refitting reproduces the stored reference trajectory. Parameter micro-perturbations at scale 10^{-4} retain trajectory correlations above 0.99999. The large changes observed after full support or source perturbations therefore cannot be explained by ordinary numerical instability.

Table 9: Correlation with the reference Reddit trajectory under selected perturbations.

Perturbation	Fixed model	Full refit
Remove lexical presence	.953	.415
Remove lexical intensity	.975	.162
Remove TextBlob	.991	.473
Remove VADER	.992	.964
Remove Mental-RoBERTa DSC	.894	.318
Start analysis in 2012	.998	.034
Start analysis in 2014	.999	.030
Start analysis in 2016	1.000	.001
High removed fraction	.777	.371
Low full-text fraction	.983	.033
Low post volume	.985	.032
Joint quality tails	.765	.003

C.2 Composite sensitivity diagnostic

The composite diagnostic variance is

$$V_t^{\text{stab}} = V_t^{\text{post}} + V_t^{\text{id}} + V_t^{\text{support}} + V_t^{\text{block}}, \tag{22}$$

with posterior-variance fraction

$$\rho_t = \frac{V_t^{\text{post}}}{V_t^{\text{stab}}}. \tag{23}$$

Because these components are not derived as independent posterior random variables, V_t^{stab} is a sensitivity construction rather than a Bayesian posterior variance. Likewise, ρ_t is a variance-composition statistic rather than a probability of correctness.

Figures 5 and 6 show the time-varying decomposition of the composite diagnostic and the aggregate mean contribution of each variance component.

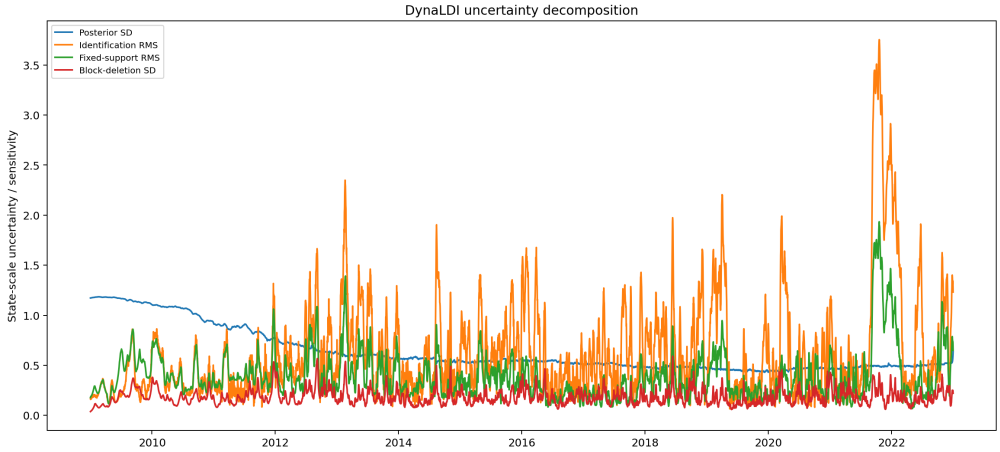


Figure 5: Time-varying decomposition of the composite diagnostic variance into posterior, identification, fixed-support, and block-deletion components.

Across the Reddit trajectory, median ρ_t is 0.479. Mean variance shares are approximately 45.96% posterior, 36.05% identification, 13.88% fixed-support, and 4.10% block-deletion sensitivity.

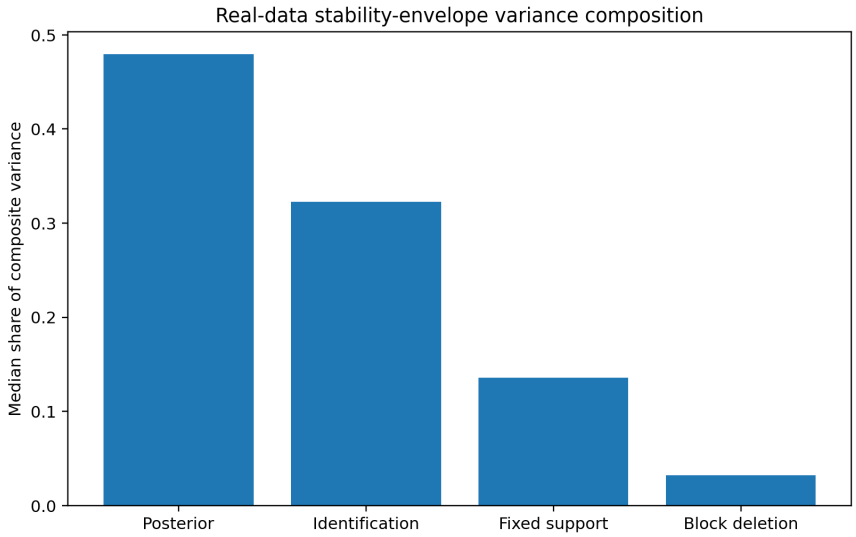


Figure 6: Mean contribution of each component to the composite diagnostic variance across the Reddit trajectory.

C.3 Late-2021 sensitivity

The standardized reference annual state for 2021 is -1.212. The full-refit perturbation consensus is approximately -0.095, while removing the highest-removal support under the fixed measurement mapping yields an annual mean of approximately -0.283. This divergence motivates the paper's decision not to attach a historical or clinical interpretation to the late-2021 trough. Figure 7 shows the annual reference trajectory together with the fixed-model and full-refit perturbation consensuses underlying this comparison.

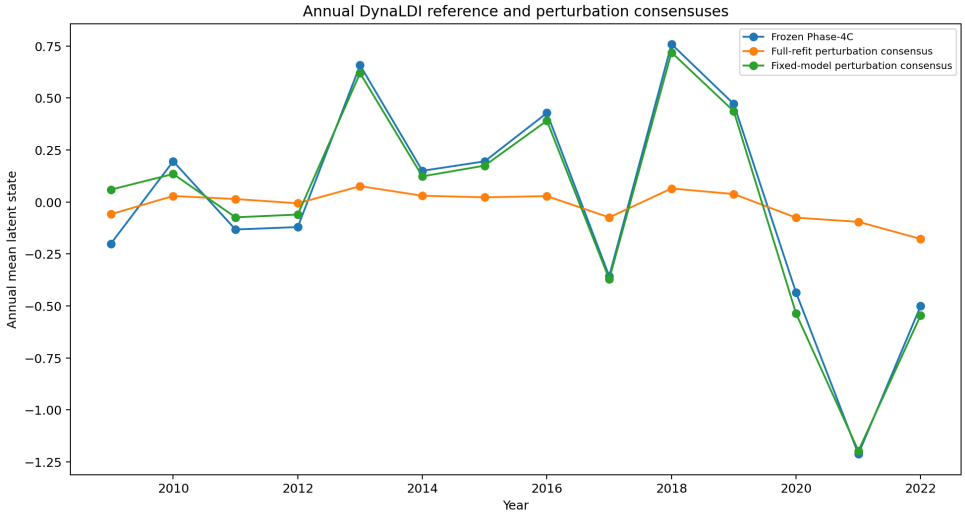


Figure 7: Annual DynaLDI reference state and perturbation consensuses. The pronounced 2021 disagreement between the reference/fixed-model estimates and the full-refit consensus illustrates sensitivity to measurement-model re-identification.

D INDEPENDENT VALIDATION OF THE STABILITY DIAGNOSTICS

This section independently evaluates whether the proposed stability diagnostics behave as intended when the latent state is known. It tests pointwise error ranking, aggregate model-health discrimination, sensitivity-envelope coverage, and detection of severe identification failures across a separate 16-scenario benchmark. A separate known-truth benchmark evaluates the stability diagnostics themselves rather than selecting the estimator. It contains

$$16 \text{ scenarios} \times 24 \text{ replicates} = 384$$

previously unseen sequences using replicate identifiers 120–143. Because these are known-truth simulations, the latent state is observed by the evaluator, allowing each stability diagnostic to be compared directly with true state-estimation error and empirical interval coverage.

The original state-recovery stresses are augmented with weak-factor, dual-factor, loading-rotation, latent-confounded-reliability, removal-process-shift, and dominant-source variance-collapse scenarios. The estimator, perturbation families, stability definitions, and confirmatory criteria are fixed before evaluation.

The validation audit applies source leave-one-out and nuisance leave-one-out perturbations, data-quality support masking at the pre-specified 10% tail, temporal-window restrictions, and random contiguous block deletion. Block-deletion diagnostics use 14- and 30-day windows targeting 10% deletion, with 20 replicates for each block length. These perturbation families and their diagnostic definitions are held fixed throughout the validation benchmark.

The eight pre-specified confirmatory criteria and their observed outcomes are reported in Table 10.

Table 10: Pre-specified stability-validation criteria.

Criterion	Observed	Result
Median Spearman: $1 - \rho_t$ vs. $\lvert e_t \rvert$	.036	Fail
Median Spearman: identification RMS vs. $\lvert e_t \rvert$	.068	Fail
Median high/low-instability RMSE ratio	1.100	Fail
Scenarios with higher RMSE in high-instability group	1.000	Pass
Severe-regime coverage improvement	.128	Pass
Matched full-envelope coverage	.981	Pass
Median envelope/posterior width ratio	1.538	Pass
Pooled high/low-instability MAE ratio	1.134	Fail

Four of eight confirmatory criteria pass, while all four benchmark-integrity checks pass. The failed criteria are retained rather than redefining the diagnostic after observing the benchmark.

The four integrity checks verify negligible variance-floor activation under matched conditions, at least 90% posterior coverage under the matched benchmark, at least 90% posterior coverage in the clean scenario, and substantial variance-floor activation under the dominant-source variance-collapse stress. The observed values are 0.000, 0.9347, 0.9338, and 0.8333, respectively, and all satisfy their pre-specified thresholds.

Figures 8 and 9 summarize the scenario-level coverage comparison and the relationship between the stability diagnostics and true estimation error after aggregation.

The ordinary posterior SD is the strongest tested pointwise error indicator, with median Spearman correlation approximately 0.112 with absolute state error. Identification RMS has median pointwise correlation 0.068, and $1 - \rho_t$ only 0.036. The stability quantities therefore do not provide strong day-level error ranking.

Their behavior is more informative after aggregation. Mean identification RMS correlates with true RMSE at Spearman 0.434 across complete simulated datasets and 0.759 across the 16 scenario means; the scenario-level association with MAE is 0.850.

The full sensitivity envelope is conservative when the measurement model is approximately correct but more useful under severe misspecification. Posterior-only coverage is 0.9347 for matched and near-matched conditions and 0.7157 under severe misspecification. The full envelope gives 0.9811 and 0.8435, respectively.

The dominant-source variance-collapse scenario provides an additional degeneracy test. Variance-floor activation is detected with AUROC 1.000 using either mean instability fraction or maximum

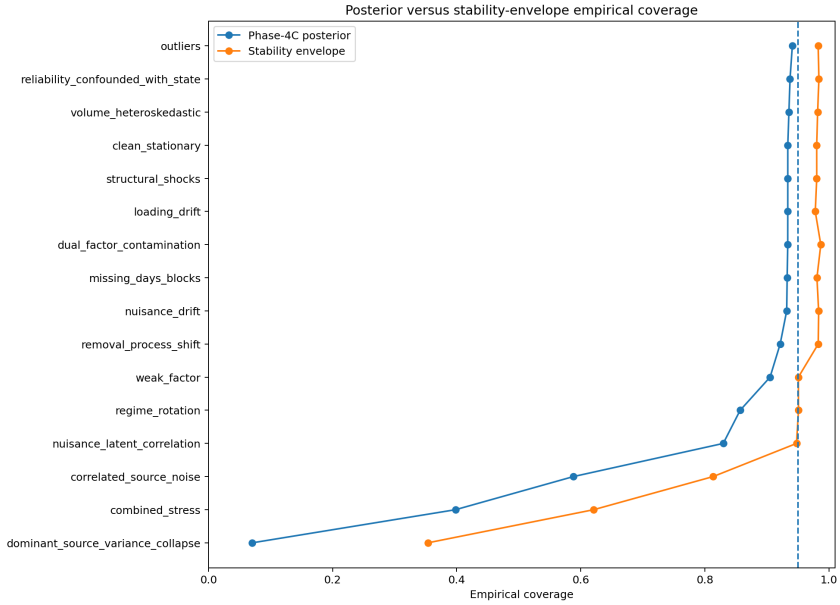


Figure 8: Independent known-truth validation: posterior-only versus full-envelope empirical coverage across validation scenarios.

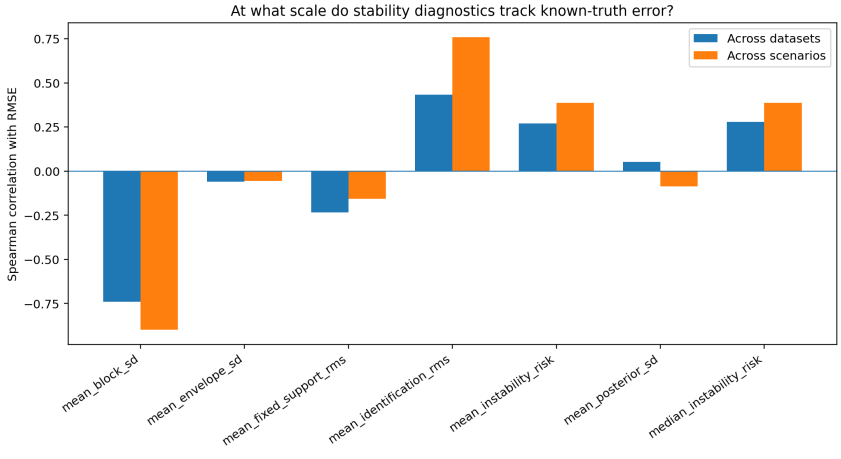


Figure 9: Independent known-truth validation: association between stability diagnostics and true estimation error after aggregation to dataset and scenario scales.

source information share. However, coverage remains poor even after envelope expansion, demonstrating that successful detection of a model-health failure does not imply that the resulting state estimate has been corrected.

E External Temporal Convergence

This section evaluates ecological temporal convergence between the fixed Reddit trajectories and five independently sourced external series. It documents temporal alignment, preprocessing and bootstrap procedures, then reports primary lag-zero comparisons and trend-robust sensitivity results.

E.1 Protocol and preprocessing integrity

All external comparisons use the already fitted Reddit trajectory. DynaLDI is not refit or retuned for any target, Equal-Z is evaluated on every target, and lag zero remains the primary alignment.

Daily Reddit values are averaged over the temporal intervals represented by each target. Because 2009 contains only 162 observed Reddit days, 2010 onward is used as the primary period for the annual and monthly targets; 2009 is retained only for sensitivity analyses.

The external comparisons are:

1. CDC Household Pulse Survey depressive-symptom estimates: 52 survey periods during 2020–2022 (National Center for Health Statistics, 2026);
2. NSDUH adult major depressive episode estimates: 10 annual observations during 2010–2019 (Substance Abuse and Mental Health Services Administration, 2020);
3. Worldwide Google Trends search interest for the literal query `depression`: 156 monthly observations during 2010–2022 (Google, 2026);
4. U.S. Google Trends search interest for the same literal query: 156 monthly observations during 2010–2022 (Google, 2026);
5. NVSS U.S. age-adjusted suicide mortality: 13 annual observations during 2010–2022 (Garnett & Curtin, 2024).

Pearson and Spearman correlations measure contemporaneous association over the aligned target periods. Standardized RMSE (zRMSE) is the root-mean-square difference after independently standardizing the Reddit trajectory and the external target over the aligned comparison period. Directional agreement is the fraction of consecutive non-tied changes for which the Reddit trajectory and external target move in the same direction.

The Google Trends definition uses the literal English Search term `depression`, Health category, and Web Search. The Worldwide series avoids the U.S.-only geographic restriction but remains language-biased. The Reddit corpus itself is not geographically restricted.

Before association analysis, the external preprocessing pipeline passed all seven raw-input integrity checks and all six target-level QA checks.

Temporal uncertainty is evaluated with paired block resampling. CDC HPS uses 2,000 circular moving-block replicates with blocks of four survey periods and base seed 9092026. NSDUH and NVSS use 10,000 paired circular moving-block replicates with three-year blocks and seed 42. Both Google Trends analyses use 10,000 paired moving-block replicates with 12-month blocks and seed 42. Within every bootstrap replicate, the same sampled indices are applied to DynaLDI, Equal-Z, and the target.

E.2 Primary and trend-robust results

Primary lag-zero results for all five external targets are summarized in Table 11.

For CDC HPS, DynaLDI gives bootstrap median Pearson 0.303 with 95% interval $[-0.023, 0.536]$ and median Spearman 0.243 with interval $[-0.087, 0.535]$. Because both correlation intervals include zero, the CDC comparison is interpreted as suggestive ecological convergence. The corresponding standardized temporal alignment is shown in Figure 10.

Table 11: Primary lag-zero external comparisons. Entries are DynaLDI / Equal-Z. Lower zRMSE is better.

Target	Pearson	Spearman	zRMSE	Directional agreement
CDC HPS	.318 / .187	.279 / -.093	1.168 / 1.275	.469 / .429
NSDUH	.230 / -.822	.289 / -.843	1.177 / 1.811	.625 / .000
Google Worldwide	.333 / -.139	.360 / -.156	1.151 / 1.504	.545 / .490
Google U.S.	.247 / -.282	.251 / -.327	1.223 / 1.596	.538 / .469
NVSS	-.249 / -.799	-.177 / -.781	1.518 / 1.823	.636 / .455

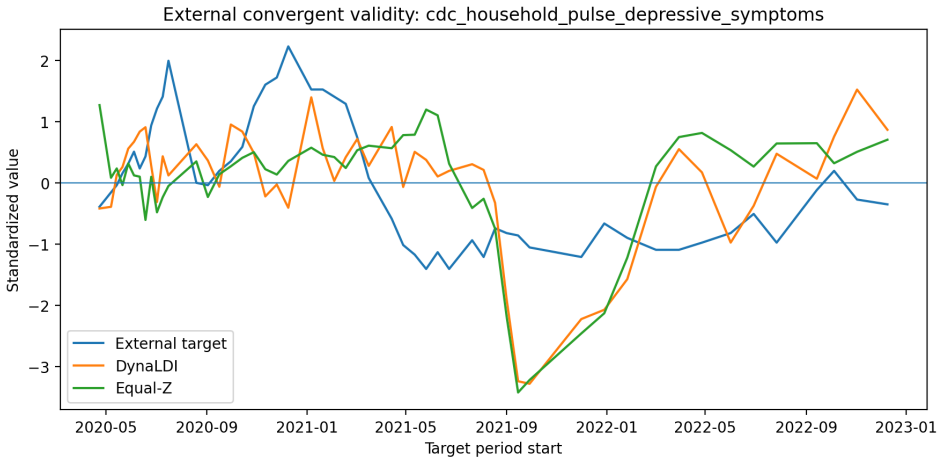


Figure 10: Standardized CDC HPS, DynaLDI, and Equal-Z temporal alignment across the 52 primary survey periods.

For NSDUH, primary DynaLDI Pearson and Spearman correlations are 0.230 and 0.289. After linear detrending they become -0.031 and -0.055, respectively, showing that the positive lag-zero association is primarily a long-horizon level/trend relationship rather than year-to-year co-movement. The annual standardized alignment is shown in Figure 11.

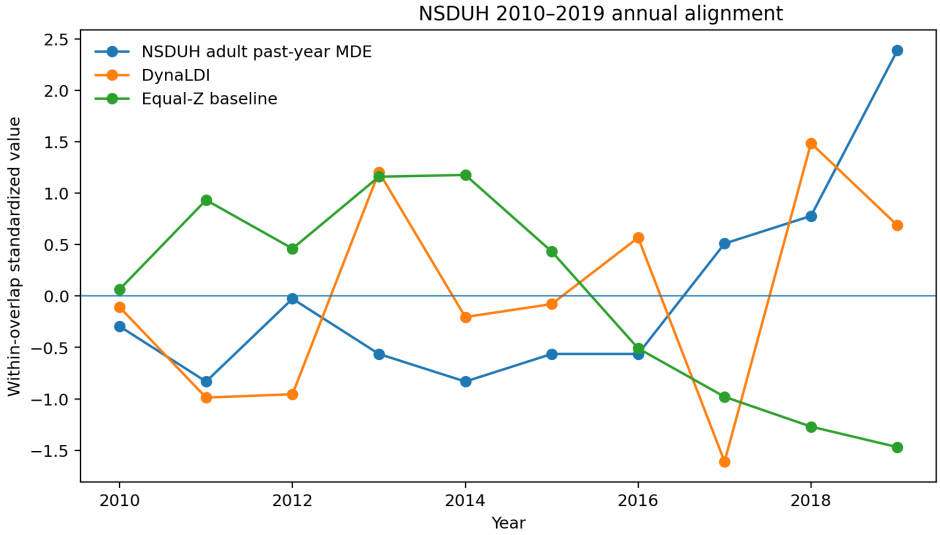


Figure 11: Annual standardized NSDUH, DynaLDI, and Equal-Z alignment over the 2010–2019 primary comparison period.

Worldwide Google Trends provides the clearest trend-robust comparison. The 95% paired-bootstrap intervals for the DynaLDI advantage over Equal-Z are $[0.224, 0.962]$ for Pearson, $[0.247, 0.945]$ for Spearman, and $[0.175, 0.696]$ for Equal-Z minus DynaLDI zRMSE. DynaLDI remains positively associated after first differencing (Pearson 0.360, Spearman 0.276) and linear detrending (Pearson 0.386, Spearman 0.389). The corresponding 12-month rolling visualization is shown in Figure 12.

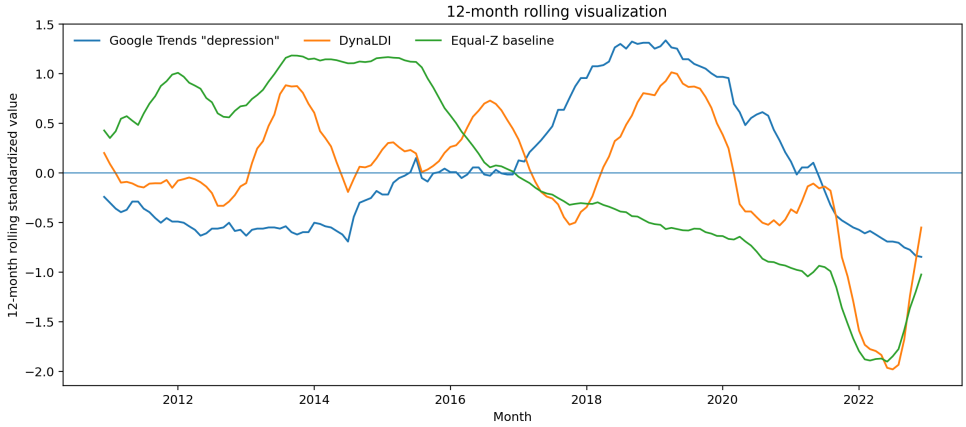


Figure 12: Worldwide Google Trends and Reddit 12-month rolling standardized alignment over the 2010–2022 primary comparison period. The rolling curves are shown for visualization only.

For U.S. Google Trends, detrended DynaLDI correlations remain positive (Pearson 0.381, Spearman 0.349), while first-difference correlations are 0.217 and 0.209. The corresponding 12-month rolling visualization is shown in Figure 13.

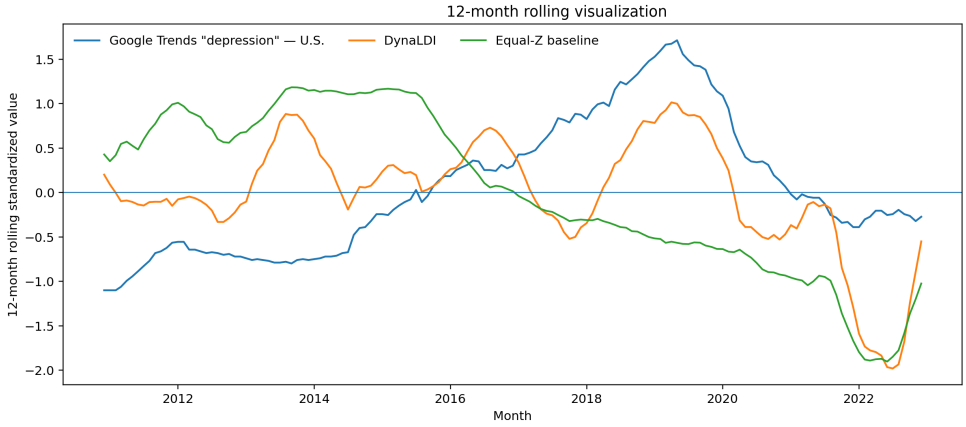


Figure 13: U.S. Google Trends and Reddit 12-month rolling standardized alignment over the 2010–2022 primary comparison period. The rolling curves are shown for visualization only.

NVSS provides a deliberate construct-boundary result. DynaLDI remains negatively associated with age-adjusted suicide mortality at lag zero (Pearson -0.249, Spearman -0.177). Equal-Z is more strongly negative, but the DynaLDI association itself is not positive. The result is therefore not presented as convergent validation. The annual standardized alignment is shown in Figure 14.

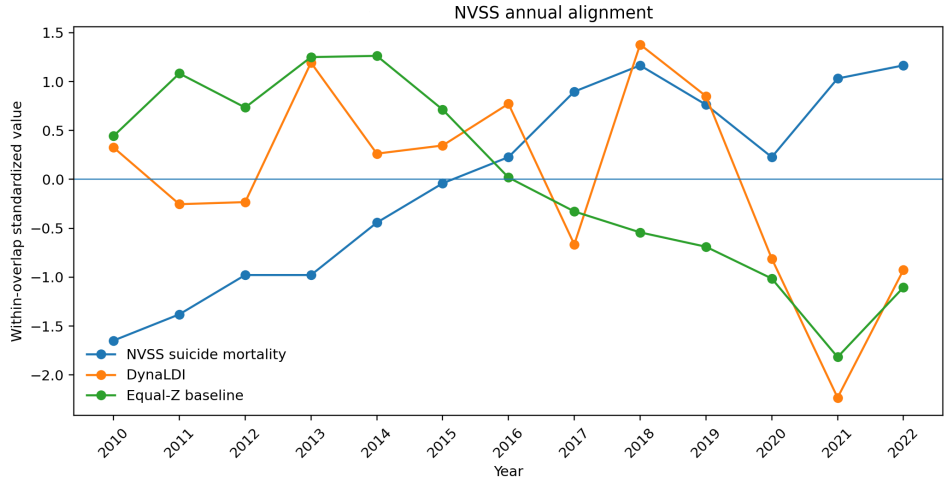


Figure 14: Annual standardized NVSS suicide mortality, DynaLDI, and Equal-Z alignment over the 2010–2022 primary comparison period.

All external comparisons operate at an ecological temporal level. They do not convert the Reddit trajectory into an estimate of depression prevalence, clinical diagnosis, symptom severity, or individual risk.

F Computational Scaling and Numerical Reproducibility

This section evaluates the computational behavior and numerical reproducibility of the scalar DynaLDI state-space kernel separately from statistical state recovery. It reports empirical runtime and memory scaling, deterministic and numerical-precision checks, and a lightweight likelihood experiment for (ϕ, q). The computational benchmark evaluates the scalar DynaLDI state-space kernel over all 30 combinations

$$T \in \{500, 1000, 2500, 5000, 10000, 20000\}, \tag{24}$$

and

$$M \in \{3, 5, 10, 20, 50\}. \tag{25}$$

The benchmark measures sufficient-statistic construction, Kalman filtering, RTS smoothing, total runtime, peak resident memory, deterministic rerun agreement, and float32-versus-float64 numerical agreement. It evaluates the state-space kernel rather than the complete joint DynaLDI optimizer.

For diagonal conditional measurement covariance and a scalar latent state, construction of the measurement sufficient statistics costs $O(TM)$, whereas filtering and smoothing after this reduction cost $O(T)$. Empirically, the fitted log–log scaling relation is

$$\log(\text{runtime}) = c + 0.9995 \log T + 0.0315 \log M, \tag{26}$$

with

$$R^2 = 0.9997. \tag{27}$$

Figure 15 summarizes the empirical runtime and peak-memory scaling across the full 30-configuration benchmark.

As shown in Figure 15, at the largest tested configuration, $T = 20{,}000$ and $M = 50$, total kernel runtime is approximately 0.082 seconds and peak resident memory is approximately 238 MiB. Approximately 0.013 seconds is spent constructing sufficient statistics, 0.044 seconds in filtering, and 0.025 seconds in RTS smoothing.

All 30 deterministic reruns reproduce the smoothed trajectory exactly:

$$\max_t \left| \widehat{d}_t^{(1)} - \widehat{d}_t^{(2)} \right| = 0. \tag{28}$$

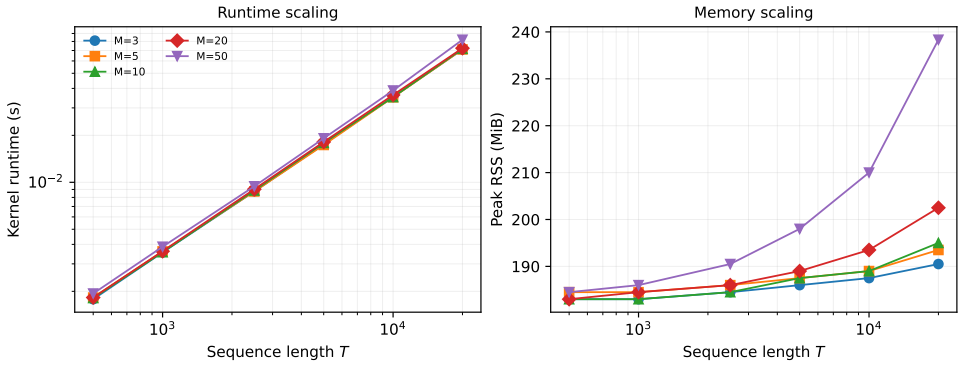


Figure 15: Empirical runtime and memory scaling of the DynaLDI state-space kernel across the 30 tested sequence-length and source-count configurations. Absolute timings are hardware dependent.

All 30 float32-input checks also pass. The worst float32-versus-float64 smoothed-state correlation is effectively one, and the maximum absolute state difference is approximately

$$1.48 \times 10^{-8}. \tag{29}$$

A separate lightweight likelihood experiment re-estimates only (ϕ, q) while holding the data-generating measurement parameters fixed. All ten pre-specified optimization configurations converge and reproduce. At $T = 500$, the optimizer reaches the lower ϕ bound, illustrating that numerical convergence does not imply strong statistical identification in a short sequence. At $T = 20{,}000$ and $M = 5$,

$$\hat{\phi} = 0.9548, \qquad \hat{q} = 0.001673, \tag{30}$$

close to the generating values 0.955532 and 0.001653. Numerical reproducibility and statistical identification are therefore treated as distinct properties.

G Reproducibility Controls and Interpretation Boundaries

The experimental workflow separates estimator development from confirmatory evaluation. The final DynaLDI configuration was fixed before the 480-sequence state-recovery benchmark and was not tuned using those confirmatory results. The perturbation families, diagnostic definitions, and acceptance criteria used in the independent 384-sequence stability benchmark were likewise specified before that benchmark was evaluated and were not redefined after observing its outcomes. External target definitions, primary analysis periods, and lag-zero alignment were specified before association analysis, and no external series was used to fit or retune DynaLDI. Computational validation was conducted separately from the statistical state-recovery and stability experiments.

The following interpretation boundaries apply throughout the paper:

1. **Latent-state meaning.** DynaLDI estimates a relative latent state jointly supported by the analyzed depression-related discourse measurements. It is not an estimate of depression prevalence, symptom severity, diagnosis, population mental-health burden, or individual risk.
2. **Weak-measurement interpretation.** The five input signals are imperfect linguistic measurements rather than clinical observations. In particular, Mental-RoBERTa DSC is a depression–suicidality contrast score derived from a subreddit-based classification task; it is not a probability of depression, suicidality, diagnosis, or individual clinical status.
3. **Filtered-state interpretation.** The filtered state uses observations only through time t, but the measurement-model parameters are estimated globally from the complete dataset. It should therefore be interpreted as observation-one-sided inference conditional on the fitted measurement model, not as a fully prospective estimate.

4. **Conditional posterior uncertainty.** The Kalman/RTS state variance quantifies uncertainty conditional on the fitted measurement specification. It does not account for uncertainty about whether a different measurement decomposition would be identified after changing sources, nuisance specification, temporal support, or data-quality support.
5. **Sensitivity-envelope interpretation.** The composite sensitivity envelope combines conditional posterior, identification, fixed-support, and block-deletion components as a model-health diagnostic. Because these components are not derived as independent posterior random variables, the envelope is not interpreted as a calibrated Bayesian credible interval or frequentist confidence interval.
6. **Posterior-variance fraction.** The quantity ρ_t describes the fraction of the composite diagnostic variance attributable to the conditional posterior. It is not a probability that the inferred state is correct, a confidence score, or a calibrated day-level reliability measure.
7. **External comparisons.** Survey, search-behavior, and mortality series are used only to assess ecological temporal convergence. They are not treated as ground truth for the Reddit latent state and do not establish individual-level or clinical validity. U.S.-specific targets additionally have a geographic mismatch with the unrestricted Reddit corpus, while the Worldwide Google Trends comparison remains limited by its literal English-language query.
8. **Late-2021 sensitivity.** The late-2021 trough is not assigned a substantive historical or clinical interpretation because its magnitude changes materially under measurement-model re-identification and removal-related observation support restrictions.
9. **Computational interpretation.** The reported scaling and reproducibility results characterize the tested scalar state-space kernel, not the complete joint DynaLDI optimizer. Numerical convergence and deterministic reproducibility should therefore be distinguished from statistical identification.