MOBILE-HUMAN-BENCH: A UNIFIED BENCHMARK FOR GENERALIZABLE MOBILE HUMAN SENSING

Anonymous authors
Paper under double-blind review

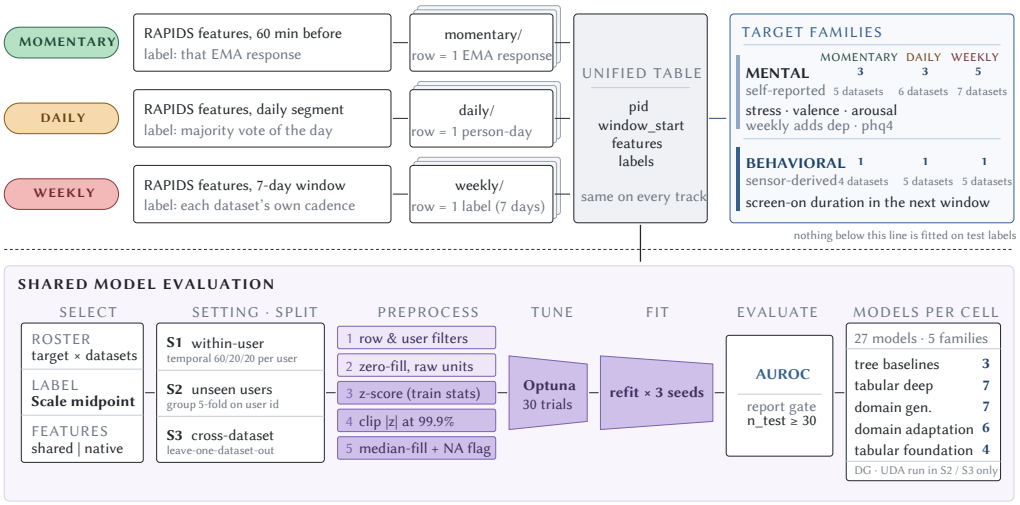


Figure 1: **Overview of Mobile-Human-Bench.** Seven in-the-wild mobile and wearable sensing collections are harmonized across momentary, daily, and weekly prediction tracks. A unified pipeline evaluates within-user temporal prediction (**S1**), generalization to unseen users (**S2**), and cross-dataset transfer (**S3**) across classical, deep tabular, DG, UDA, and tabular foundation models.

ABSTRACT

Mobile sensing enables modeling of human states and behaviors, yet progress is difficult to compare across studies because datasets differ in sensing modalities, targets, temporal resolution, preprocessing, and evaluation. We introduce **Mobile-Human-Bench**, a unified benchmark for evaluating generalization across time, users, and independently collected studies. It harmonizes seven longitudinal studies spanning **1,751 participants across nine countries** and yields **295,162 labeled feature rows** after benchmark construction. The benchmark operates at the structured sensing-feature level, supports momentary, daily, and weekly prediction, and covers affective, mental-health, and behavioral targets. We evaluate classical machine learning, deep tabular models, domain generalization, unsupervised domain adaptation, and tabular foundation models under within-user temporal, unseen-user, and cross-dataset evaluation. Results show that model rankings and apparent gains change substantially across generalization settings, targets, and timescales. Tabular foundation models are strong in within-user settings but do not retain a systematic advantage for unseen users or independently collected studies, while standard domain generalization and adaptation provide limited gains over matched controls. Mobile-Human-Bench provides a reproducible testbed for identifying which modeling advances remain robust across users, studies, outcomes, and timescales. Resources are available at `https://anonymous.4open.science/r/Mobile-Human-Benchmark-B3C5`.

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.

1 INTRODUCTION

Smartphones and wearable devices enable continuous, in-the-wild measurement of behavioral, contextual, and physiological signals, supporting prediction of mental health and well-being (Mohr et al., 2017; Nepal et al., 2024; Pratap et al., 2022), affective states (Meegahapola et al., 2022; Ruensuk et al., 2020; Park et al., 2020), and everyday behaviors (Busso et al., 2025; Nepal et al., 2024). Yet naturalistic sensing introduces substantial variability across individuals, populations, time, and collection conditions (Meegahapola et al., 2022; Xu et al., 2023; Singh et al., 2025; Zhang et al., 2026b). Independently collected studies further differ in devices, sensing availability, targets, temporal resolution, feature representation, preprocessing, and evaluation protocols. Consequently, improvements reported within one study may not persist across users or datasets. This limitation is particularly important for pretrained sensing and tabular models, whose gains may depend strongly on the users, datasets, and evaluation protocols under which they are assessed (Abbaspourazad et al., 2024; Erturk et al., 2025; Xu et al., 2025b).

Existing benchmark-style efforts address parts of this problem. DiversityOne studies geographic generalization across countries (Meegahapola et al., 2022; Busso et al., 2025); GLOBEM evaluates longitudinal and cross-dataset transfer across repeated annual collections (Xu et al., 2022; 2023); and EmoPhone separates within-user, cross-user, and cross-wave evaluation within a coordinated multi-wave study (Zhang et al., 2026b). FEEL is the closest multi-dataset precedent, harmonizing 19 independently collected datasets for physiological emotion recognition (Singh et al., 2025). However, these efforts do not provide a common benchmark spanning independently collected behavioral and contextual mobile sensing studies across multiple targets, prediction timescales, and generalization settings.

A central challenge is the input representation. Mobile sensing combines regularly sampled signals with sparse and asynchronous event streams, such as calls, messages, application launches, and screen interactions; structured health data more broadly often mixes continuous measurements with irregular discrete events (Jing et al., 2026). Across independently collected studies, these modalities further differ in availability, sampling, logging schemas, and temporal coverage, making a common raw or sequential representation difficult without substantial modality-specific alignment (Liu et al., 2026). Such a representation would often restrict evaluation to the small subset of signals consistently collected across studies, excluding much of the behavioral and contextual information central to mobile sensing. We therefore harmonize studies at the *structured-feature level*, consistent with recent work on behavioral and sensor-derived structured representations (Erturk et al., 2025; Felizzi, 2026).

Mobile-Human-Bench. We introduce **Mobile-Human-Bench**, a unified benchmark over seven longitudinal smartphone and wearable sensing studies (Figure 1). Mobile-Human-Bench standardizes sensing-derived representations, target construction, temporal aggregation, preprocessing, data splitting, model selection, and evaluation while preserving source-dataset boundaries. It supports **momentary**, **daily**, and **weekly** prediction across affective, mental-health, and behavioral targets, using structured sensing features constructed or aligned through the **Reproducible Analysis Pipeline for Data Streams (RAPIDS)** (Vega et al., 2021).

The benchmark defines three complementary generalization settings: **S1**, within-user temporal prediction from earlier to later observations of the same participant; **S2**, within-dataset generalization to unseen participants; and **S3**, cross-dataset generalization to an independently collected held-out study (Xu et al., 2023; Singh et al., 2025; Zhang et al., 2026b). Across these settings, we compare classical machine-learning baselines, deep tabular models, domain generalization (DG), unsupervised domain adaptation (UDA), and tabular foundation models. This design tests whether modeling advantages persist as evaluation moves from familiar users to unseen users and ultimately to new studies, and whether conclusions change with prediction timescale.

Research questions. We organize the benchmark around three research questions: **RQ1:** How does predictive performance change across S1, S2, and S3? **RQ2:** Which modeling approaches remain robust across datasets and generalization settings? **RQ3:** How do benchmark conclusions change across momentary, daily, and weekly prediction?

Table 1: **Positioning of Mobile-Human-Bench against representative sensing benchmarks.** We compare GLOBEM (Xu et al., 2022; 2023), EmoPhone (Zhang et al., 2026b), and FEEL (Singh et al., 2025). T, U, and D denote temporal/longitudinal, cross-user, and independent cross-dataset evaluation, respectively.

Work	Organization	Devices	Data type	Targets	T	U	D	Timescales
GLOBEM	Coordinated annual datasets	Smartphone + wearable	Behavioral / contextual features	Behavior / mental health	✓	✓	–	Longer-horizon
EmoPhone	Three coordinated waves	Smartphone + wearable	Behavioral / contextual + physiological features	Affect	✓	✓	–	Momentary
FEEL	19 independent datasets	Wearable / physiological sensors	Physiological features	Emotion	Mixed	Mixed	✓	Heterogeneous
Ours	**7 independent studies**	**Smartphone + wearable**	**Behavioral / contextual + physiological features**	**Affect / mental health / behavior**	✓	✓	✓	**Momentary / daily / weekly**

Contributions. Our contributions are threefold. **C1: A unified structured-data benchmark for in-the-wild mobile human sensing.** We harmonize seven independently collected longitudinal datasets under common feature, target, preprocessing, and evaluation protocols while preserving dataset boundaries. **C2: A broad evaluation across generalization settings, model families, and timescales.** We benchmark S1–S3 across classical, deep tabular, DG, UDA, and tabular foundation methods at momentary, daily, and weekly resolutions where supported. **C3: An empirical characterization of robustness.** We quantify how predictive performance and conclusions about model families change across temporal, interpersonal, and study-level shifts, and which findings remain stable across targets and timescales.

Together, Mobile-Human-Bench provides a common testbed for evaluating whether advances in mobile human sensing remain reliable beyond the dataset and population in which they were developed.

2 Related Work

2.1 Mobile Human Sensing Datasets and Benchmarks

Smartphones and wearables support in-the-wild prediction of affective states, mental-health outcomes, and behavioral quantities over timescales ranging from momentary self-reports to daily and longer-horizon outcomes (Harari & Gosling, 2023; Mohr et al., 2017). Representative datasets include StudentLife (Wang et al., 2014), the College Experience Study (Nepal et al., 2024), DiversityOne (Busso et al., 2025), GLOBEM (Xu et al., 2022), K-EmoPhone (Kang et al., 2023), and EmoPhone (Zhang et al., 2026b). These resources differ substantially in population, collection duration, sensing configuration, feature availability, label density, target definition, and temporal resolution, making conclusions difficult to compare across studies.

Prior benchmark-style efforts broaden evaluation beyond conventional within-dataset prediction. DiversityOne studies geographic and population generalization (Meegahapola et al., 2022; Busso et al., 2025); GLOBEM evaluates longitudinal and cross-dataset generalization across coordinated annual collections (Xu et al., 2022; 2023); and EmoPhone separates within-user, cross-user, and cross-wave evaluation within a coordinated multi-wave affect-sensing study (Zhang et al., 2026b). FEEL provides the closest precedent for independent multi-dataset evaluation, harmonizing 19 datasets for physiological emotion recognition from EDA and PPG (Singh et al., 2025). As summarized in Table 1, Mobile-Human-Bench instead spans heterogeneous structured smartphone and wearable sensing across independently collected studies, multiple prediction timescales, and temporal, cross-user, and cross-dataset evaluation.

2.2 Structured-Data and Foundation Models

Mobile sensing pipelines commonly transform heterogeneous sensor streams and behavioral events into structured features (Vega et al., 2021; Xu et al., 2022). More generally, health data may combine regularly sampled measurements with sparse and irregular events, motivating structured representations for heterogeneous inputs (Jing et al., 2026). Strong tabular baselines include tree ensembles and deep architectures such as TabNet, TabTransformer, SAINT, and FT-Transformer (Arik & Pfister, 2021; Huang et al., 2020; Somepalli et al., 2021; Gorishniy et al., 2021), while TabPFN and related models introduce pretrained priors for tabular prediction (Hollmann et al., 2023).

In parallel, sensing foundation models have been pretrained on large physiological, motion, and behavioral sensing data (Abbaspourazad et al., 2024; Narayanswamy et al., 2025; Pillai et al., 2025; Xu et al., 2025a; Erturk et al., 2025). Many operate within relatively standardized modalities or sensing ecosystems, whereas independent mobile studies combine heterogeneous modalities, event streams, logging schemas, and temporal coverage. Mobile-Human-Bench therefore evaluates general-purpose tabular models on semantically aligned structured sensing features, while retaining dataset-native representations where appropriate.

2.3 Generalization under Distribution Shift

Mobile sensing models face multiple forms of distribution shift: temporal changes within participants, interpersonal differences across users (Zhang et al., 2026a), population and geographic shifts (Meegahapola et al., 2022; Nanchen et al., 2023), and study-level shifts across independently designed datasets. The latter may jointly change recruitment, sensing protocols, feature availability, target construction, and collection conditions, producing a compound shift beyond any single population or temporal difference.

Domain generalization (DG) targets unseen domains using source data alone, whereas unsupervised domain adaptation (UDA) additionally uses unlabeled target-domain observations (Zhou et al., 2023; Wilson & Cook, 2020). TableShift benchmarks DG under real-world tabular shifts (Gardner et al., 2023); GLOBEM provides a mobile-sensing precedent for DG (Xu et al., 2023); and M3BAT introduces UDA for multimodal mobile sensing (Meegahapola et al., 2024). However, classical, deep-tabular, DG, UDA, and tabular foundation methods have not been systematically compared within one mobile-sensing benchmark spanning temporal, cross-user, and independent cross-dataset shifts across multiple prediction timescales.

3 Benchmark Characterization

Mobile-Human-Bench harmonizes seven longitudinal mobile and wearable sensing studies while preserving their study boundaries. We characterize the benchmark along three axes: datasets, prediction targets and timescales, and feature representations. Full dataset cards, label mappings, sensing inventories, and preprocessing details are provided in Appendix A.

3.1 Dataset Suite and Heterogeneity

Table 2 summarizes the seven benchmark datasets. They span approximately a decade of data collection (2013–2023), nine countries, study durations from about one week to multiple years, and label schedules ranging from dense in-situ reports to daily and weekly assessments. The datasets also differ in population, geography, devices, sensing availability, study protocol, and label instruments. We treat these differences jointly as *study-level heterogeneity*; cross-dataset performance should therefore not be attributed to any single source of shift.

3.2 Prediction Targets and Temporal Units

Prediction tasks are defined jointly by *target* and *timescale*. **Stress, valence, and arousal** form the main shared affective targets. Longer-horizon mental-health outcomes, including PHQ-4 and a BDI-II-derived depression indicator, are retained where supported, while **next-window screen-on duration** provides a sensor-derived behavioral target. Only datasets with compatible measurements participate in a given target roster; dataset-specific targets are restricted to within-dataset evaluation.

Table 2: **Benchmark source datasets.** Years, locations, participant counts, durations, and label cadence summarize the source releases before benchmark-specific filtering and harmonization. Post-harmonization sample sizes, target rosters, and benchmark temporal constructions are reported in Appendix A.

Collection	Year(s)	Location	N	Duration	Label cadence
StudentLife	2013	USA	49	10 wk	momentary
K-EmoPhone	2019	S. Korea	77	∼1 wk	∼10/day
EmoPhone	2020–22	S. Korea	297	∼4 wk/wave	∼4–8/day
DiversityOne	2020–21	8 countries	782	∼4 wk/site	∼30 min
DeepStress	2023	S. Korea	24	∼7 wk	∼11/day
CES	2017–22	USA	216	∼4 yr	daily
GLOBEM	2018–21	USA	497	∼3 mo/wave	weekly

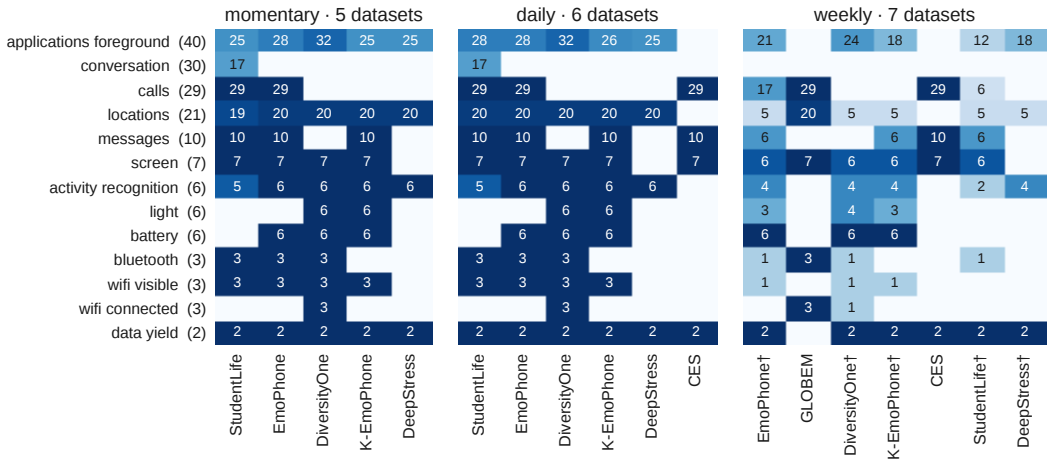


Figure 2: Canonical sensing-feature availability across datasets and prediction timescales. Differences in coverage determine the Shared feature spaces available for cross-dataset evaluation.

Self-reported affective scales are oriented consistently and binarized at the midpoint of their declared response range. PHQ-4 is binarized at the published moderate-or-worse threshold of 6, the released depression indicator is retained as provided, and screen-on duration is binarized by participant-specific median. Complete mappings are provided in Appendix A.

We construct three temporal tracks: **momentary**, using sensing from the preceding 60 minutes for an in-situ report; **daily**, pairing person-day features with a target from the same day; and **weekly**, using 7- or 14-day sensing histories paired with the target cadence supported by each dataset. Daily affective results are thus concurrent inference, not forecasting, as a day's sensing can extend past its prompts (Appendix A.6); screen-on labels always follow the features.

3.3 Common Feature Representation

We harmonize sensing data at the structured-feature level using RAPIDS (Vega et al., 2021). The canonical schema covers application use, communication, context, location, screen interaction, activity, Wi-Fi, light, battery, and Bluetooth sensing; compatible pre-derived features are mapped to the same semantics. Feature availability nevertheless varies across datasets and timescales (Figure 2).

We evaluate two feature spaces. **Native** retains benchmark-valid features available within each dataset, including dataset-specific sensing features. **Shared** begins from a structural schema defined for each target–timescale roster as the intersection of semantically compatible features available across all participating datasets. This schema is fixed across settings and transfer directions within the roster and is determined from feature definitions and availability only. After an evaluation split is formed, value-dependent validity filtering is fitted on the training side, so the final model inputs

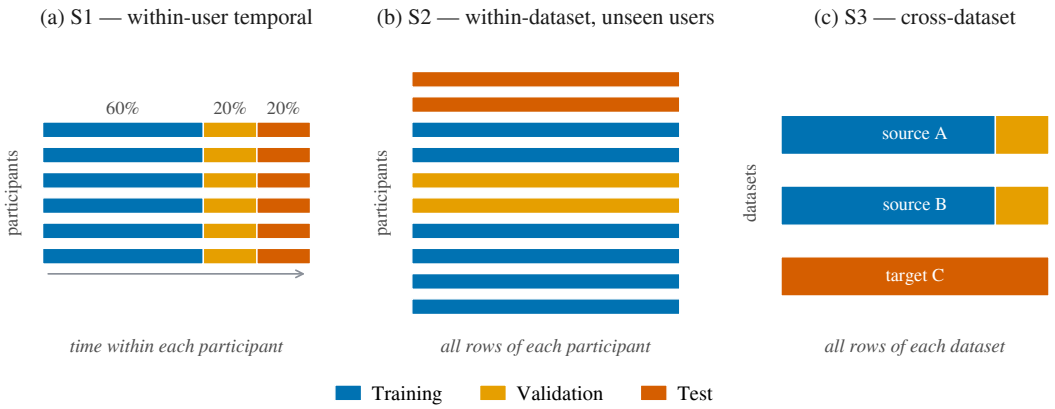


Figure 3: **Three benchmark evaluation settings.** S1 uses chronological within-user 60/20/20 splits; S2 holds out participants using grouped 5-fold evaluation; S3 holds out an entire target dataset with source-only validation.

may be a subset of the structural Shared schema. S1 and S2 support both Native and Shared representations, whereas S3 requires Shared features for cross-dataset transfer. No wearable feature is currently shared canonically across the benchmark, so Shared schemas are handset-derived. Detailed feature counts, validity filtering, missingness rules, and preprocessing are reported in Appendix A.

4 Benchmark Evaluation and Findings

We evaluate Mobile-Human-Bench under three deployment settings and use the resulting benchmark to study how conclusions change across distribution shifts, model families, prediction timescales, and outcomes. Full split definitions, model configurations, hyperparameter spaces, preprocessing details, and per-dataset results are provided in the Appendix.

4.1 Evaluation Protocol

Generalization settings. The benchmark defines three complementary settings. **S1** evaluates within-user temporal prediction: each participant's observations are ordered chronologically and divided into 60/20/20 training, validation, and test portions. For momentary prediction, we restrict each participant to the first 30 days following the reference protocol; daily and weekly tracks use the available history. **S2** evaluates generalization to unseen participants from the same study using participant-grouped 5-fold evaluation, with validation participants drawn only from the training side. **S3** evaluates transfer to an independently collected held-out dataset. Validation is restricted to source participants and target labels are never used for preprocessing, model selection, or fitting. Supervised and DG methods receive source data only, whereas UDA may additionally observe unlabeled target features. Cross-dataset evaluation uses only features shared by the participating datasets. S1, S2, and S3 therefore expose temporal, interpersonal, and compound study-level shifts, respectively.

Models and controlled comparisons. We evaluate 27 models from five families: conventional machine-learning baselines, deep tabular models, DG, UDA, and tabular foundation models. Baseline, deep-tabular, and foundation models are evaluated across all three settings; DG and UDA are evaluated in S2 and S3. An *evaluation cell* is defined by its setting, target, feature space, and dataset or transfer direction. Within a cell, applicable models receive identical splits, features, labels, preprocessing, and evaluation rows. Tunable methods receive the same 30-trial Optuna budget with validation AUROC for selection, whereas pretrained tabular foundation models are evaluated without task-specific hyperparameter search. DG and UDA methods are additionally compared with matched source-only controls using the same backbone where applicable.

Metrics and statistics. We run three seeds for every cell; S2 additionally uses five participant-disjoint folds. AUROC is the primary metric, pooled over all test rows, and results are reported only

Table 3: **Family-level benchmark summary on the shared affective targets.** Mean test AUROC across applicable stress, valence, and arousal evaluation cells. DG and UDA are undefined for S1. **Bold** marks the highest family mean within each timescale–setting column. Complete model- and target-level results are reported in Appendix C.

	Momentary			Daily			Weekly		
Family	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Baseline	.646	.540	.510	.587	.528	.516	.571	.505	**.499**
Deep Tabular	.566	.540	.517	.531	.530	.516	.541	.512	.497
DG	–	.542	**.519**	–	.529	.517	–	**.516**	.496
UDA	–	.544	**.519**	–	**.539**	**.519**	–	.514	.496
Tabular FM	**.693**	**.547**	.508	**.643**	.529	.500	**.580**	.496	.497

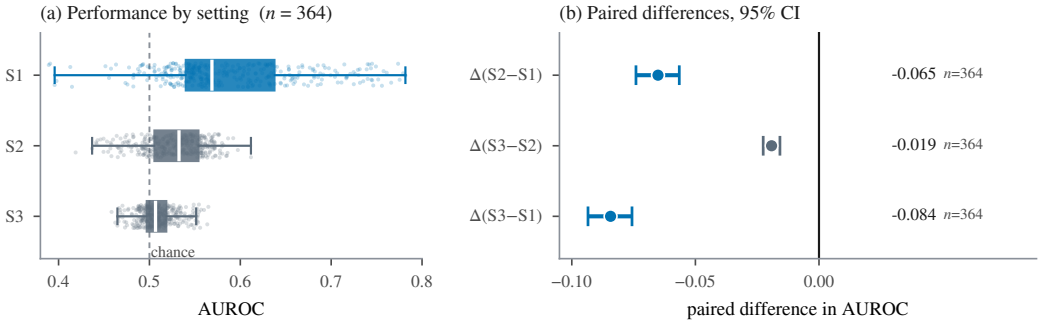


Figure 4: **Generalization across evaluation settings.** (a) AUROC distributions over the 364 matched affective configurations. (b) Paired differences between settings with 95% bootstrap confidence intervals. Performance declines from S1 to S2 and further from S2 to S3.

when the test split contains both classes and at least 30 rows. Inferential comparisons are paired at the evaluation-cell level using two-sided Wilcoxon signed-rank tests, bootstrap confidence intervals for paired mean differences, and Holm–Bonferroni correction. For null effects, we additionally test practical equivalence within a ± 0.01 AUROC margin. Implementation and statistical details are provided in Appendix B.

4.2 Generalization Across Evaluation Settings

Table 3 summarizes family-level performance, while Figure 4 compares 364 matched configurations over the shared affective targets—stress, valence, and arousal—holding target, dataset, timescale, feature space, and model fixed. Complete model- and target-level results are provided in Appendix C.

Affective prediction degrades from S1 to S3. Mean AUROC decreases from 0.5929 in S1 to 0.5278 in S2 and 0.5086 in S3. The paired differences are -0.0651 for S2–S1, -0.0192 for S3–S2, and -0.0843 for S3–S1 (Figure 4); all remain significant after Holm correction. Performance also compresses around chance: the AUROC IQR falls from 0.0981 to 0.0490 and 0.0221 across S1–S3.

4.3 Model Families Under Distribution Shift

Figure 5 shows a strong interaction between model family and evaluation setting. Family differences are substantial under S1 but compress sharply when participants or entire studies are held out.

Shift-aware objectives do not improve over matched controls. Across 24 matched DG/UDA method–setting comparisons, none yields a statistically significant positive improvement over its source-only control: 21 are practically equivalent within ± 0.01 AUROC and three show small significant decreases. UDA therefore provides no measurable advantage despite observing unlabeled target features.

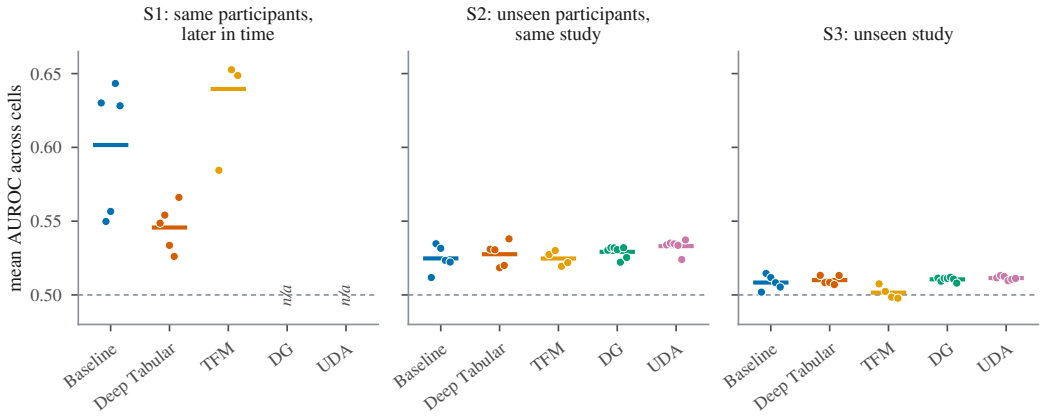


Figure 5: **Model-family performance across evaluation settings.** Points show individual model means over the shared affective evaluation cells and horizontal bars show family means. Differences between model families are pronounced in S1 but substantially compressed in S2 and S3. DG and UDA are undefined in S1.

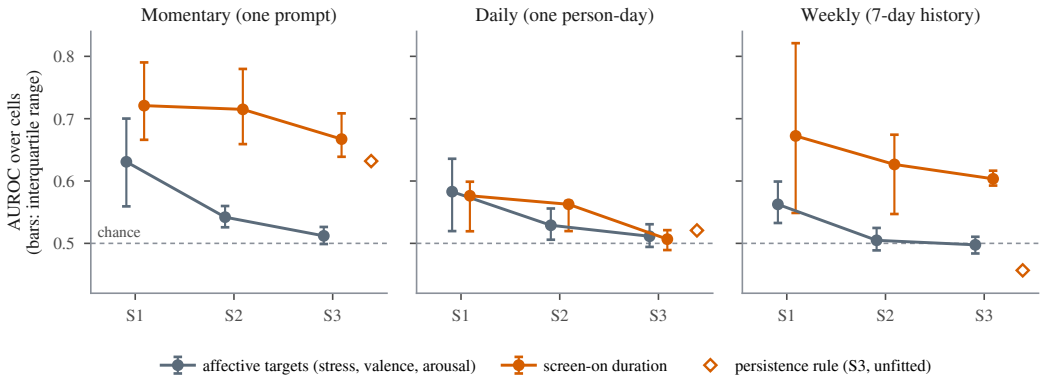


Figure 6: **Outcome-dependent generalization across prediction timescales.** Lines show mean AUROC across evaluation cells for the shared affective targets (stress, valence, and arousal) and screen-on duration; error bars denote interquartile ranges. The affective targets consistently degrade toward chance from S1 to S3, whereas screen-on duration retains stronger performance at selected timescales. Diamonds show the unfitted persistence rule under S3, illustrating how much behavioral predictability can be explained by simple temporal carry-over. The affective and screen-on series summarize their respective eligible cell rosters and should not be interpreted as matched target-level comparisons.

Foundation-model gains do not persist beyond S1. Tabular foundation models have the highest S1 family mean (0.6398) and win 44 of 54 shared-affective S1 cells, but the five family means converge to 0.5247–0.5331 in S2 and 0.5015–0.5115 in S3. Strong within-user performance therefore does not imply cross-user or cross-study generalization.

4.4 Outcome and Timescale Dependence

Because the tracks differ in sensing history, row construction, coverage, cadence, and temporal alignment, we compare generalization patterns within timescale rather than raw performance across timescales. Figure 7 shows that momentary, daily, and weekly prediction all reproduce the S1 $>$ S2 $>$ S3 pattern. The smaller weekly S2–S3 gap reflects a floor effect: weekly affective prediction is already near chance under S2.

The behavioral target provides a contrasting transfer pattern. As shown in Figure 6, screen-on duration retains substantially higher cross-study performance at the momentary and weekly

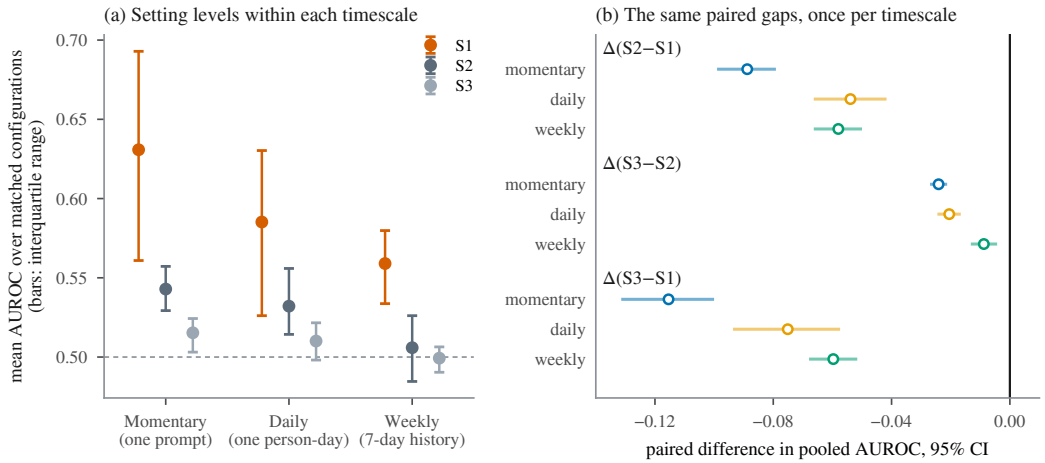


Figure 7: **Setting effects by prediction timescale.** (a) Mean AUROC and interquartile range within each timescale. (b) Paired setting differences with 95% confidence intervals. Comparisons are made within timescale because the three tracks differ in datasets, row construction, and target alignment.

timescales than the affective targets. However, the persistence baseline shows that part of this apparent robustness reflects simple carry-over rather than learned cross-study structure. The effect of distribution shift therefore depends jointly on outcome and temporal resolution.

Timescale is therefore part of the task definition, and conclusions should be reported jointly over setting, target, and timescale.

4.5 Implications and Limitations

Mobile-Human-Bench separates performance under familiar users from robustness under deployment shift. Models that are clearly separated under S1 become difficult to distinguish under S2, and especially under S3; the screen-on results show that this compression is outcome-dependent, although much of what survives is carry-over. Claims of generalizable mobile sensing should therefore report participant and dataset overlap, temporal construction, and target-domain access.

The benchmark nevertheless has several limitations. It currently covers seven datasets, with substantial representation from student and young-adult populations and a limited set of affective, mental-health, and behavioral targets. Our structured-feature representation is a deliberate trade-off: it enables comparison across heterogeneous continuous signals and sparse behavioral event streams whose modalities, sampling, and logging schemas differ substantially across studies, but it does not preserve all fine-grained temporal dynamics or raw-signal information. Cross-dataset evaluation further requires semantically compatible shared features, excluding some study-specific and wearable-derived information. Dataset coverage also differs across targets and timescales, and S3 captures compound study-level heterogeneity rather than an isolated population shift. Future extensions should broaden populations and outcomes and complement the structured benchmark with raw or sequential evaluation where sufficiently compatible sensing streams are available.

5 Conclusion

We introduced **Mobile-Human-Bench**, a harmonized structured sensing-feature benchmark for in-the-wild mobile human sensing across seven longitudinal studies. Performance and model-family conclusions change substantially across within-user, unseen-user, and cross-dataset evaluation. Tabular foundation models are strong under within-user temporal prediction but do not retain a systematic advantage for unseen users or independently collected studies, while standard DG and UDA provide limited gains over matched controls. **Mobile-Human-Bench provides a reproducible testbed for identifying which modeling advances persist across users, studies, outcomes, and timescales**, complementing foundation-model efforts built on more unified sensing ecosystems.

AI USAGE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing portions of the manuscript text and with programming assistance, including code completion, debugging, and refactoring. We did not use generative AI tools to independently design experiments, analyze results, or derive scientific conclusions. All AI-assisted text was reviewed and revised by the authors, and AI-assisted code was inspected, tested, and validated against the intended experimental procedures and outputs. We take responsibility for the final content of this work, including all text, code, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work performs secondary analysis of previously collected mobile and wearable sensing datasets and follows the access and usage conditions of each source resource. To reduce privacy risks, the released benchmark contains only processed, extracted sensing features and task labels required for benchmark evaluation; we do not redistribute raw sensor streams or personally identifying information. In particular, privacy-sensitive raw data such as precise GPS traces and other direct or device-level identifiers are not included in the benchmark release. The benchmark is intended for research on robust and generalizable mobile sensing, and all released artifacts are limited to the information necessary to reproduce the reported evaluation.

REPRODUCIBILITY STATEMENT

Section 4.1 defines the evaluation settings, model families, metrics, and common experimental protocol; Table 19 records which models report in which setting and timescale. Appendix A documents dataset construction, harmonization, target definitions, and eligibility criteria; Appendix B provides complete preprocessing, hyperparameter search, training, and implementation details; and Appendix C reports the complete benchmark results. The seven source datasets are third-party resources released under their own access conditions, and raw sensor streams, raw survey responses, and device-level identifiers are not redistributed. We release the harmonization pipeline, feature and target configurations, experiment configurations, split indices, per-cell result matrices, and the processed feature tables and task labels used in every experiment. Every reported number can therefore be regenerated, and new models can be evaluated on identical cells without access to the source releases. Anonymous source code, configurations, and data are available at `https://anonymous.4open.science/r/Mobile-Human-Benchmark-B3C5`.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *International Conference on Learning Representations*, 2024.

Sercan Ö. Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021. doi: 10.1609/aaai.v35i8.16826.

Matteo Busso, Andrea Bontempelli, Leonardo Javier Malcotti, Lakmal Meegahapola, Peter Kun, Shyam Diwakar, Chaitanya Nutakki, Marcelo Dario Rodas Britez, Hao Xu, Donglei Song, Salvador Ruiz Correa, Andrea-Rebeca Mendoza-Lara, George Gaskell, Sally Stares, Miriam Bidoglia, Amarsanaa Ganbold, Altangerel Chagnaa, Luca Cernuzzi, Alethia Hume, Ronald Chenu-Abente, Roy Alia Asiku, Ivan Kayongo, Daniel Gatica-Perez, Amalia de Götzen, Ivano Bison, and Fausto Giunchiglia. Diversityone: A multi-country smartphone sensor dataset for everyday life behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(1), 2025. doi: 10.1145/3712289.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J. Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15516–15541, 2025.

Federico Felizzi. Accelerometry-derived digital biomarkers for cardiometabolic risk: A population-representative tabular benchmark with uncertainty quantification. *arXiv preprint arXiv:2606.30702*, 2026.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. In *Advances in Neural Information Processing Systems*, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Gabriella M. Harari and Samuel D. Gosling. Understanding behaviours in context using mobile sensing. *Nature Reviews Psychology*, 2:767–779, 2023. doi: 10.1038/s44159-023-00235-3.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. TabTransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Zilin Jing, Vincent Jeanselme, Yuta Kobayashi, Simon A. Lee, Chao Pang, Aparajita Kashyap, Yanwei Li, Xinzhuo Jiang, and Shalmali Joshi. One loss to rule them all: Marked time-to-event for structured ehr foundation models. *arXiv preprint arXiv:2602.00541*, 2026.

Soowon Kang, Woohyeok Choi, Cheul Young Park, Narae Cha, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Heepyung Kim, Yong Jeong, and Uichin Lee. K-EmoPhone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels. *Scientific Data*, 10(1):351, 2023. doi: 10.1038/s41597-023-02248-2.

Yuxuan Liu, Joshua Placidi, Jinpei Han, Alfred John Balston, Marek Rei, and A. Aldo Faisal. Autoregressive ehr foundation models with multimodal inputs. *arXiv preprint arXiv:2607.22264*, 2026.

Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, José Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Gunel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. Generalization and personalization of mobile sensing-based mood inference models: An analysis of college students in eight countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4), 2022. doi: 10.1145/3569483.

Lakmal Meegahapola, Hamza Hassoune, and Daniel Gatica-Perez. M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2), 2024. doi: 10.1145/3659591.

David C. Mohr, Mi Zhang, and Stephen M. Schueller. Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13: 23–47, 2017. doi: 10.1146/annurev-clinpsy-032816-044949.

Alexandre Nanchen, Lakmal Meegahapola, William Droz, and Daniel Gatica-Perez. Keep sensors in check: Disentangling country-level generalization issues in mobile sensor-based models with diversity scores. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, pp. 217–228. Association for Computing Machinery, 2023. doi: 10.1145/3600211.3604688.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam A. Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94b25992757a549470c8f8dfe73d8df6-Abstract-Conference.html.

Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. Capturing the college experience: A four-year mobile sensing study of mental health, resilience and behavior of college students during the pandemic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1), 2024. doi: 10.1145/3643501.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020. doi: 10.1038/s41597-020-00630-y.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7871f51d332c8661cc729539d7987b21-Abstract-Conference.html.

Abhishek Pratap, Ava Homiar, Luke Waninger, Calvin Herd, Christine Suver, Joshua Volponi, Joaquin A. Anguera, and Pat Areán. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. *Scientific Data*, 9(1):522, 2022. doi: 10.1038/s41597-022-01633-7.

Mintra Ruensuk, Eunyong Cheon, Hwajung Hong, and Ian Oakley. How do you feel online: Exploiting smartphone sensors to detect transitory emotions during social media use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4), 2020. doi: 10.1145/3432223.

Pragya Singh, Ankush Gupta, Somay Jalan, Mohan Kumar, and Pushpendra Singh. Feel: Quantifying heterogeneity in physiological signals for generalizable emotion recognition. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025) Datasets and Benchmarks Track*, 2025.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Julio Vega, Meng Li, Kwesi Aguillera, Nikunj Goel, Echhit Joshi, Kirtiraj Khandekar, Krina C. Durica, Abhineeth R. Kunta, and Carissa A. Low. Reproducible analysis pipeline for data streams: Open-source software to process data collected with mobile devices. *Frontiers in Digital Health*, 3, 2021. doi: 10.3389/fdgth.2021.769823.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella M. Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14, 2014. doi: 10.1145/2632048.2632054.

Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):1–46, 2020. doi: 10.1145/3400066.

Maxwell A. Xu, Jaya Narain, Gregory Darnell, Haraldur T. Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Andres Fineman, Karthik Jayaraman Raghuram, James Matthew Rehg, and Shirley You Ren. RelCon: Relative contrastive learning for a motion foundation model for wearable data. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/83eb339ed42297658fa24b5cec939285-Abstract-Conference.html.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*,

volume 35, pp. 24655–24692. Curran Associates, Inc., 2022. doi: 10.52202/068431-1790. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9c7e8a0821dfcb58a9a83cbd37cc8131-Paper-Datasets_and_Benchmarks.pdf.

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4), January 2023. doi: 10.1145/3569485. URL https://doi.org/10.1145/3569485.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized foundation models struggle to beat supervised baselines. In *International Conference on Learning Representations*, 2025b.

Panyu Zhang, Minseo Park, Tomiris Ismatzoda, Azizbek Mustafakulov, Uzair Ahmed, Otabek Najimov, Jumabek Alikhanov, Surjya Ghosh, and Uichin Lee. CrossShift: Quantifying interpersonal differences in mobile sensing for mental health. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(3), September 2026a. ISSN 2474-9567. doi: 10.1145/3831954. URL https://doi.org/10.1145/3831954.

Panyu Zhang, Minseo Park, Soowon Kang, Tomiris Ismatzoda, Azizbek Mustafakulov, Otabek Najimov, Woohyeok Choi, Jumabek Alikhanov, Surjya Ghosh, and Uichin Lee. EmoPhone: A multi-wave dataset for in-the-wild mobile and wearable affect sensing. *arXiv preprint arXiv:2609.16581*, 2026b. doi: 10.48550/arXiv.2609.16581.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.

A Benchmark Datasets, Targets, and Features Documentation

A.1 Benchmark scope and the resources admitted

A *resource* is one entry of the benchmark roster and the unit every evaluation setting treats as a dataset; a *wave* is a sub-collection released as its own table. Five of the seven resources are a single wave; EmoPhone pools three annual waves and GLOBEM four, so the roster is assembled from twelve released tables. Pooling is not concatenation: EmoPhone's waves reuse participant identifiers for different people, so identifiers are namespaced by wave before any split is computed, each wave passes its own quality filter first, and the pooled feature space is the intersection of the waves rather than their union, because carrying the union would leave a block of light-sensor columns missing in exactly the pattern of a wave indicator.

Table 4: **Benchmark resources as built.** Counted from the table each track loaded, at the finest timescale the resource enters (*At*: M/D/W). *Targets* counts this resource's harmonized targets and names its core ones: str stress, val valence, aro arousal, phq4 the PHQ-4 score, dep the BDI-II-derived depression score. *Part.* counts GLOBEM's person-waves, not distinct individuals. *Rows* counts rows that carry a target label after the label–feature join, up to 9.6% below the labelled count of Table 9.

Resource	Comp.	Tracks	Targets	Part.	Rows	At
K-EmoPhone	single	M/D/W	6 (str, val, aro)	77	5,485	M
EmoPhone	3 waves	M/D/W	17 (str, val, aro)	297	53,100	M
DeepStress	single	M/D/W	1 (str)	24	12,791	M
StudentLife	single	M/D/W	1 (str)	49	2,099	M
DiversityOne	single	M/D/W	1 (val)	384	179,846	M
CES	single	D/W	1 (str)	217	34,954	D
GLOBEM	4 waves	W	2 (phq4, dep)	703	6,887	W

The distinction between a *coordinated multi-wave resource* and an *independently collected study* is what S3 rests on. EmoPhone's and GLOBEM's waves were collected by one team to one protocol, so pooling them recombines a single design rather than merging two populations; the other five are independent studies, and S3 reads a held-out resource as an unseen study, which is only meaningful where a resource boundary is also a protocol and population boundary. Four of the seven come from one research group and share a logging schema family, so seven resources are not seven independent samples of the task.

Five criteria decide admission, stated as preconditions required by the harmonization process rather than as quality thresholds applied in advance: *C1*, a target in the harmonized roster; *C2*, in-situ rather than laboratory collection, so that a row represents a moment in everyday life rather than a condition in an experimental protocol; *C3*, phone sensing that yields the shared feature layer, either as raw streams that RAPIDS can process or as natively provided RAPIDS-equivalent aggregates; *C4*, data accessibility; and *C5*, at least one timescale–target cell with both classes present and at least 30 test observations, following the reporting rules introduced in Section 4.1. C3 determines a resource's eligible *tracks* more often than its admission. GLOBEM, for example, provides aggregate features rather than raw sensing streams and is admitted under the second clause of C3, entering the benchmark only at the weekly timescale.

One candidate resource was screened formally and not adopted. OpenMHC (MyHeartCounts) fails C3, because its nineteen HealthKit channels contain none of location, screen, applications, activity recognition, Wi-Fi, Bluetooth, calls, messages, light or battery, and under the intersection rule of Appendix A.12 admitting it would collapse a shared space from 43 usable columns to steps alone; it fails C1, because its 169-variable registry carries no PHQ, GAD, PSS, BDI or CES-D and its depression and anxiety items are single retrospective wellbeing questions, which would make an instrument join into a construct join; and it fails C5 on scale, with 11,894 users against 1,942 for the rest of the roster combined. The cost of that exclusion is the roster's main population limitation: every admitted resource is a student or small young-adult cohort, so the population axis rests on DiversityOne's multi-site contrast alone.

Table 5: **Resource-admission criteria and exclusions.** The table summarizes the benchmark inclusion criteria, the candidate resource formally evaluated but not adopted, and exclusions applied within resources that otherwise remain in the benchmark.

Item	Criterion or exclusion	Rationale
A. Resource-inclusion criteria		
C1	Target availability	The resource contains at least one item mapped in Section 3 to stress, valence, arousal, PHQ-4, or the BDI-II-derived depression score. The item must be administered in situ or according to a declared survey schedule.
C2	In-situ collection	Labels are collected during everyday activity. Consequently, each observation represents a moment in everyday life rather than an experimental or laboratory condition.
C3	Compatible phone sensing	The resource provides participant-level passive sensing streams that RAPIDS can process or supplies natively extracted RAPIDS-equivalent aggregate features.
C4	Data accessibility	The resource is publicly available or accessible under a data-use agreement held by the project.
C5	Reportable evaluation cell	After applying the evaluation rules introduced in Section 4.1 and detailed in Appendix A.8, at least one timescale–target cell remains with both classes present and at least 30 test observations.
B. Evaluated candidate not adopted		
OpenMHC (MyHeartCounts)	C1, C3, and C5	**Sensor incompatibility (C3):** Its 19 HealthKit channels include steps, distance, flights climbed, heart rate, active energy, sleep, and workout indicators. However, they exclude the principal sensing families used by the benchmark, including location, screen, applications, activity recognition, Wi-Fi, Bluetooth, calls, messages, light, and battery. Under the shared-space intersection rule, admitting OpenMHC would reduce the 43-feature shared space to steps alone. **Target incompatibility (C1):** Its 169-variable registry contains no PHQ, GAD, PSS, BDI, or CES-D instrument. Its depression and anxiety variables are individual retrospective well-being questions rather than compatible instruments. **Scale imbalance (C5):** Its 11,894 users substantially exceed the 1,942 participant records in the remaining resource roster.
C. Exclusions within admitted resources		
GLOBEM	Momentary track	The release provides daily feature aggregates rather than raw event streams, and its surveys are administered approximately once or twice per week. It therefore cannot support a momentary sensing window.
GLOBEM	Daily track	Although daily aggregate features are available, the target instruments are administered approximately once or twice per week. GLOBEM consequently enters the benchmark only at the weekly timescale.
CES	Momentary track	CES labels are recorded daily and contain no momentary response timestamp with which to anchor a preceding sensing window.
K-EmoPhone	14-day history block	The study lasts eight days and therefore cannot provide a complete 14-day history. Including this block would introduce 52 empty columns into the shared feature space.

Exclusions also occur *inside* an admitted resource, each with a recorded basis rather than a judgement. DiversityOne's Ulaanbaatar site was never converted, so the resource contributes seven of its eight countries and the roster's ten countries become nine in the built tables. GLOBEM's first wave leaves the PHQ-4 tasks, having administered a two-item substitute on 0–4 rather than the instrument. CES never enters the momentary track, its labels carrying no response timestamp to anchor a sensing window. DeepStress carries no screen stream at all.

A.2 Participants and the sample funnel

The seven resources contribute 1,751 participant records as built, of which GLOBEM's 703 are person-waves rather than distinct individuals, the release giving no cross-year identity linkage. A built count is not a released cohort size, and the two differ in both directions. Filtering reduces: DiversityOne's 782 released participants become 384 as built, because the Ulaanbaatar site was never converted and because the row-quality and minimum-response filters of Appendix A.13 remove participants whose sensing coverage or response count falls below the benchmark's floors before any split is computed. The person-wave convention increases: GLOBEM's 497 released individuals become 703 person-waves, one record per individual per wave they appear in, since the release gives no identifier linking a participant across its four annual collections. Two smaller differences arise in the same way. CES lists 216 participants in the release's demographics file, the figure reported in Table 2, whereas its EMA and sensing tables cover 220 identifiers, of which 217 have at least one labelled day that joins with sensing; the benchmark uses these 217. DiversityOne's weekly table retains 364 of its 384 participants, because a weekly row is anchored on a daily sensing summary for the day preceding the response, and 20 participants, each with at most two daily summaries, have no response that meets this condition. The funnel down to the cells that are actually modelled is steep in a way that is a property of the evaluation setting rather than of data quality.

Table 6: **Participant and study statistics.** The table shows the reduction from built participant records to those retained for modelling. *Retained* reports the range across a resource's eligible cells because participant eligibility is determined per cell rather than per resource. Native and shared feature spaces produce the same participant retention. S3 does not have an equivalent retention count because it transfers to an entire unseen resource. *Built* counts participants in the harmonized table of each timescale, which differs from the released cohort sizes of Table 2 as explained in Appendix A.2; GLOBEM is counted in person-waves. Demographic statistics are reported in the original publication associated with each resource.

Resource	Timescale or scope	Built	Retained S1	Retained S2
A. Overall participant-retention ranges				
K-EmoPhone	All eligible cells	77	16–77	71–77
EmoPhone	All eligible cells	297	89–297	92–297
DeepStress	All eligible cells	24	20–24	24
StudentLife	All eligible cells	49	4–46	10–47
CES	All eligible cells	217	203–212	203–213
DiversityOne	All eligible cells	384	24–347	272–348
GLOBEM	All eligible cells	703	151–692	549–703
B. Participant retention by timescale				
K-EmoPhone	Momentary	77	71–77	77
K-EmoPhone	Daily	77	16–41	76–77
K-EmoPhone	Weekly	77	49–61	71
EmoPhone	Momentary	297	89–296	92–297
EmoPhone	Daily	297	188–241	297
EmoPhone	Weekly	297	280–288	297
DeepStress	Momentary	24	24	24
DeepStress	Daily	24	20	24
DeepStress	Weekly	24	24	24

Continued on the next page

Table 6 continued from the previous page

Resource	Timescale or scope	Built	Retained S1	Retained S2
StudentLife	Momentary	49	32	35
StudentLife	Daily	49	36	47
StudentLife	Weekly	49	4	10
CES	Daily	217	205	205
CES	Weekly	217	212	213
DiversityOne	Momentary	384	219	348
DiversityOne	Daily	384	24	331
DiversityOne	Weekly	364	171	295
GLOBEM	Weekly	703	151–340	549–703

Holding out each participant's own later observations, as S1 does, requires both classes on both sides of that participant's own chronological split, and that removes far more of a cohort than holding out whole participants does: StudentLife retains between 4 and 46 of its 49 participants under S1 against 10 to 47 under S2, and DiversityOne between 24 and 347 of 384 under S1 against 272 to 348 under S2. A range is given rather than a single number because a participant drops out per cell, not per resource — the same person may be modelled for one target at one timescale and be ineligible for another.

A.3 Sensing availability

An intersection over a set of resources is bounded by the sparsest column in it, so the per-family counts below explain most of what Appendix A.12 reports. Wearable sensing is available in four source releases but is not uniformly represented in the benchmark feature spaces. K-EmoPhone's MS Band 2 and EmoPhone's Fitbit streams do not enter the built benchmark tables. In contrast, GLOBEM retains Fitbit-derived sleep and step features in its weekly table, and these features are used in its Native-space single-dataset experiments. No wearable feature is part of the canonical schema, however, so all Shared cross-dataset feature spaces are handset-derived.

Table 7: **Sensing availability across benchmark resources.** Entries indicate the number of canonical features available for each sensing stream and resource, counted on the same feature registry as Figure 2; the schema size of each stream is given in brackets, and the streams together make up the 166-feature canonical schema. "–" indicates that the sensing stream is unavailable. Counts are for the daily track, except for GLOBEM, which enters only at the weekly track. K-EP denotes K-EmoPhone, EP denotes EmoPhone, DS denotes DeepStress, SL denotes StudentLife, D1 denotes DiversityOne, and GLB denotes GLOBEM. Because a shared feature space is constructed by intersection, its size is bounded by the sparsest participating resource, as reported in Table 15.

Sensing stream	K-EP	EP	DS	SL	CES	D1	GLB
Activity recognition (6)	6	6	6	5	–	6	–
Ambient light (6)	6	–	–	–	–	6	–
App usage (40)	26	28	25	28	–	32	–
Battery (6)	6	6	–	–	–	6	–
Bluetooth (3)	–	3	–	3	–	3	3
Calls (29)	–	29	–	29	29	–	29
Conversation audio (30)	–	–	–	17	–	–	–
Data yield (2)	2	2	2	2	2	2	–
Location (21)	20	20	20	20	–	20	20
Messages (10)	10	10	–	10	10	–	–
Screen (7)	7	7	–	7	7	7	7
Wi-Fi, visible (3)	3	3	–	3	–	3	–
Wi-Fi, connected (3)	–	–	–	–	–	3	3
Total (166)	86	114	53	124	48	88	62

A.4 BENCHMARK TARGETS AND THEIR INSTRUMENTS

The benchmark harmonizes five targets across resources – three affective states and two longer-horizon mental-health instruments, and carries fifteen further targets that a single instrument supports. Each affective target has its own roster, because no target is answered by every resource: stress by K-EmoPhone, the EmoPhone waves, DeepStress and StudentLife at the momentary window and by CES at the daily one; valence by K-EmoPhone, EmoPhone and DiversityOne; arousal by K-EmoPhone and EmoPhone alone. Transfer is run once per target over that target's own roster. Screen-on duration, the behavioural reference target of Section 3.2, is constructed from the sensing stream rather than from an instrument; it has one row in Table 8, is defined in the paragraph below, and its roster is given in Table 11.

Table 8: **The original instrument behind each dataset–target pair.** What the HIGH class denotes in the source paper's own words, and the measured median gap between a participant's consecutive responses. Item, scale and harmonization are in Table 9. †Screen-on is built from the screen event stream rather than an instrument, over K-EmoPhone, EmoPhone, StudentLife, CES, DiversityOne and GLOBEM (roster in Table 11); its cadence entry gives the label-window length at the momentary, daily and weekly tracks.

Resource	Target	High denotes	Cadence
K-EmoPhone	stress	more stressed	51 min
K-EmoPhone	valence	more positive emotion	51 min
K-EmoPhone	arousal	more excited (vs. calm)	51 min
EmoPhone	stress	more stressed	77 min
EmoPhone	valence	more positive emotion	77 min
EmoPhone	arousal	more excited (vs. calm)	77 min
DeepStress	stress	more stressed	63 min
StudentLife	stress	more stressed (item recoded before splitting)	127 min
CES	stress	more stressed	5.0 days
DiversityOne	valence	more positive mood (item reversed before splitting)	30 min
GLOBEM	phq4	more anxiety and depression; cut at the published bands, …	7.0 days
GLOBEM	dep	BDI-II-derived depression, adopted verbatim from the …	7.0 days
Six resources†	screen-on	more screen-on time than the participant's median over the following window (training-slice median under S1)	15 min / 1 / 7 days

Screen-on duration. The label is the screen-on time inside a window that lies wholly after the feature window: the 15 minutes after the prompt at the momentary track, the next local day at the daily track, and the next 7-day block at the weekly track. Because the label window follows the features, the full feature space, screen columns included, stays legitimate. Episodes longer than 300 minutes are treated as a missing OFF event, as RAPIDS' own screen features treat them; StudentLife is the exception, because its logger records hour-long intervals by construction. A daily or weekly row is left unlabelled when its label window holds no sensing (at least one day, or five of seven, with a screen event), so a participant whose logging stopped is not scored as a confident zero. The label is binarized at the participant's own median. *Under S1 that median is computed on the participant's training slice only* and applied unchanged to their validation and test rows. A whole-series median would define the S1 test label partly from the test rows, and it would flip 3.5–5.3% of momentary, 7.4–13.2% of daily and 10–26% of weekly labels. Under S2 and S3 the median is computed over each participant's full series. Test participants are disjoint from training participants there, so the threshold uses no training participant's data and no test row enters training; a test participant's threshold is, however, set from their own rows. A participant whose median equals their maximum, such as a constant series, would receive an all-negative label and is left unlabelled instead. DeepStress carries no screen stream, GLOBEM enters at the weekly track only, and K-EmoPhone's eight-day study holds no row with both a 7-day history and a covered next week, so K-EmoPhone is absent at the weekly track.

DiversityOne's mood item enters the valence roster through an explicit alias rather than a judgement that mood and valence are interchangeable: it is a single five-point happy-to-sad scale, reverse coded relative to every other member, and after the correction of Appendix A.5 the benchmark's columns reproduce the study's own published labels on all 198,886 rows exactly.

The two mental-health instruments are a BDI-II-derived depression score and PHQ-4 on 0–12, both at the weekly window, both on GLOBEM, and both reporting under S1 and S2 only. That is a design decision rather than an absence of data: CES also administers PHQ-4, so a GLOBEM–CES direction is nominally available and is deliberately not run, because the weekly space the two share reduces to a handful of screen and call columns, and a direction built there would be a screen-and-calls task reported as cross-dataset mental-health transfer.

A.5 Label harmonization

A harmonized label is a threshold placed on somebody else's response scale, and the threshold is a modelling decision whichever way it falls. The benchmark fixes one rule and applies it to the declared range rather than to the observed data: every rated scale is split at its own midpoint, so that HIGH denotes the same thing everywhere — above the neutral point of the scale the participant was shown. Splitting on an observed quantile would make a cross-dataset comparison a comparison between label definitions. Every rated scale here has an odd number of points, so the midpoint is an attainable response rather than a gap between options, and the boundary decision must be explicit: the midpoint response joins HIGH, a convention pinned to DiversityOne's published one.

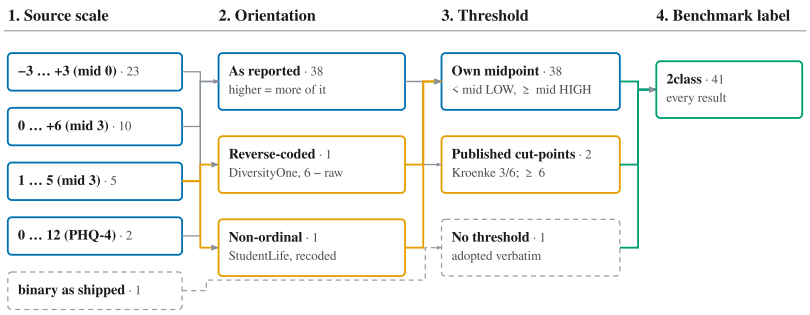


Figure 8: **The label harmonization path.** From a source response to a benchmark label: declared scale, orientation correction where the instrument needs one, midpoint threshold, two-class column. The corrections shown are StudentLife's unordered stress item and DiversityOne's reverse-coded mood item.

Two scales need a correction before the rule may be applied, and both were found by reading the source instrument rather than the data. StudentLife's stress item is not an ordered scale at all — its five options run *a little stressed*, *definitely stressed*, *stressed out*, *feeling good*, *feeling great* — so thresholding the raw codes places the two calm responses in the high-stress class, inverting 596 of 2,279 responses, 26% of the target. DiversityOne's mood item is reverse coded and is mapped to $6 - x$ first; that correction is the one place the benchmark can check itself against an external answer, and after it the generic rule reproduces that study's published labels exactly.

Table 9: **Harmonized label definitions, all targets.** One row per (resource, target) the benchmark runs; DiversityOne's mood item enters as valence. EmoPhone is one dataset here, its three waves pooled: each splits its own scale at its own midpoint, so the counts add. *Range* is the declared scale with its threshold in parentheses; stress and disturbance are named by scale length because EmoPhone's wave 4 coded those seven-point items 0..6 and waves 2 and 3 $-3..3$. *HIGH* and *mid* are percentages, *mid* the share sitting exactly on the midpoint. GLOBEM's targets bypass the harmonizer: `dep` is the release's own binary label, adopted as shipped, and `phq4` is split at ≥ 6 by the weekly builder; both are counted on the built weekly table. All reported tasks are binary.

Resource	Target	Source	Range	Fix	Rows	HIGH	mid
K-EmoPhone	arousal	arousal	-3..3 (0)	–	5,582	56.3	17
K-EmoPhone	attention	attention	-3..3 (0)	–	5,582	70.3	20
K-EmoPhone	change	change	-3..3 (0)	–	5,582	81.3	63
K-EmoPhone	disturbance	disturbance	-3..3 (0)	–	5,582	64.0	27
K-EmoPhone	stress	stress	-3..3 (0)	–	5,582	53.1	17
K-EmoPhone	valence	valence	-3..3 (0)	–	5,582	79.6	22
EmoPhone	arousal	arousal	-3..3 (0)	–	53,139	47.1	11
EmoPhone	attention	attention	-3..3 (0)	–	31,301	68.4	13
EmoPhone	change	change	-3..3 (0)	–	10,259	79.9	61
EmoPhone	disturbance	disturbance	7-pt (mid)	–	53,139	58.9	12
EmoPhone	mental	mental	-3..3 (0)	–	31,301	53.7	14
EmoPhone	stress	stress	7-pt (mid)	–	53,139	46.1	12
EmoPhone	valence	valence	-3..3 (0)	–	53,139	77.3	17
EmoPhone	arousal_change	arousalChange	-3..3 (0)	–	21,042	51.1	16
EmoPhone	valence_change	valenceChange	-3..3 (0)	–	21,042	76.0	19
EmoPhone	angry	angry	0..6 (3)	–	21,838	14.6	6
EmoPhone	anxious	anxious	0..6 (3)	–	21,838	34.3	13
EmoPhone	cheerful	cheerful	0..6 (3)	–	21,838	67.6	29
EmoPhone	content	content	0..6 (3)	–	21,838	70.2	27
EmoPhone	depressed	depressed	0..6 (3)	–	21,838	21.4	9
EmoPhone	happy	happy	0..6 (3)	–	21,838	76.1	35
EmoPhone	relaxed	relaxed	0..6 (3)	–	21,838	77.2	28
EmoPhone	sad	sad	0..6 (3)	–	21,838	16.2	8
DeepStress	stress	stressLevel	1..5 (3)	–	13,134	35.0	13
StudentLife	stress	ema_Stress_level	1..5 (3)	recode	2,279	73.8	44
CES	stress	stress	1..5 (3)	–	35,348	47.3	29
DiversityOne	valence	a6a	1..5 (3)	reverse	198,886	97.0	23
GLOBEM	dep	dep	binary	–	6,887	46.6	–
GLOBEM	phq4	phq4	0..12 (6)	–	5,046	17.2	5

PHQ-4 is the one target not split at a midpoint, using the published severity bands and binarizing at the moderate-or-worse boundary of 6. On GLOBEM this departs from that study's own published cut of 3, and the departure is adopted for comparability with CES, which ships its binary PHQ-4 at 6: at that cut CES's released labels are 10.7% positive and GLOBEM's 17.2%, where the milder cut would raise them to 37.7% and 47.8%. The cost is named rather than absorbed — LOW on PHQ-4 here includes clinically mild cases — and the raw score is retained so the alternative recomputes without a code change.

Where the neutral bucket is large, the boundary decision rather than the responses carries most of the label: 63.0% of K-EmoPhone's emotion-change responses sit exactly on the midpoint, so that target is 81.3% HIGH as constructed and would be 18.3% HIGH under the opposite convention, and results on it read as *reported a change at all* rather than *reported a positive change*. The affective targets are also not balanced and their imbalance is not uniform — DiversityOne's mood sits at 0.97–0.99 positive, K-EmoPhone and EmoPhone valence between 0.79 and 0.89, stress between 0.30 on DeepStress and 0.85 on StudentLife — so every DiversityOne result is reported beside a majority-class rate of 0.97, and a near-chance AUROC there is a property of the label rather than a failure of the model.

A.6 TEMPORAL TARGET CONSTRUCTION

Three quantities must be kept apart: the *feature history* summarized by a row, the *label window* defining its target, and the temporal relationship between them. As illustrated in Figure 9, a momentary row summarizes the 60 minutes preceding an answered ESM prompt. A daily row pairs sensing aggregated over a local calendar day with the target derived from responses collected on that same day. The daily track therefore measures concurrent day-level state inference rather than next-day forecasting. Because sensing is aggregated over the full calendar day, the feature window may extend beyond individual prompts contributing to the daily target.

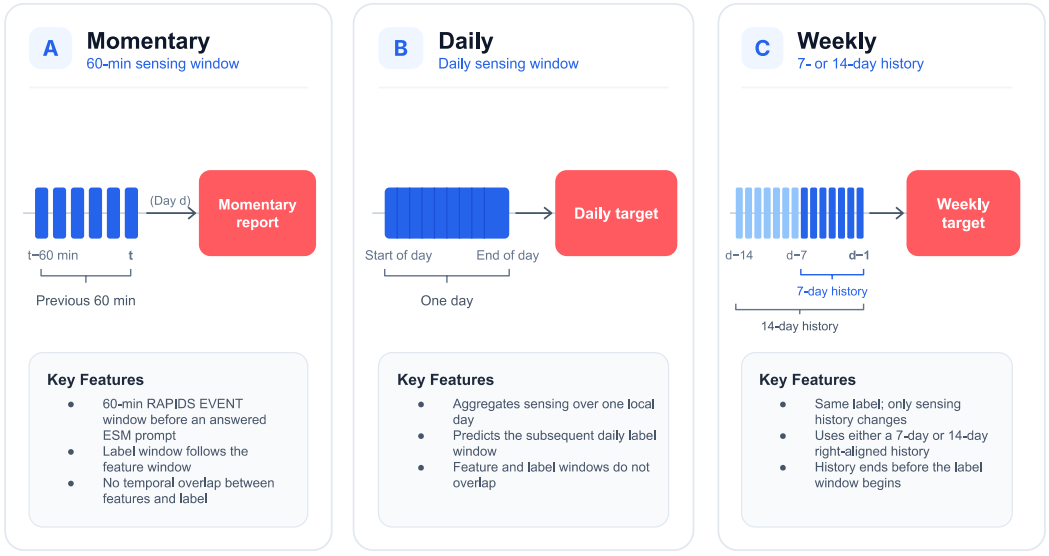


Figure 9: **Temporal target construction across the three benchmark timescales.** Momentary prediction uses sensing features from the 60 minutes preceding an answered ESM prompt. Daily inference pairs sensing aggregated over one local calendar day with a target derived from responses collected on that same day, and therefore represents concurrent day-level state inference rather than next-day forecasting. Weekly construction uses 7- or 14-day sensing histories with dataset-specific temporal alignment to the corresponding weekly target; the exact alignment is summarized in Table 10.

Table 10: **Prediction windows by timescale.** The feature history summarized by each row, the window over which its target is defined, and their temporal alignment. Temporal alignment is timescale- and dataset-specific.

Timescale	Feature window	Label	Temporal alignment	Segment
Momentary	the 60 minutes ending at the prompt	the response at the prompt	the sensing window ends at the prompt	$[t-60, t]$
Daily	one local calendar day	the majority class of answered prompts on the same day; CES uses its native daily class	same-day feature and target windows; concurrent day-level inference	one calendar day
Weekly	a 7- or 14-day sensing history	GLOBEM uses a person-week target; other datasets use their corresponding weekly target	dataset-specific alignment	7dhist, 14dhist

The weekly construction requires an important caveat. For the three shared affective targets, no resource other than CES administers a weekly instrument. Instead, the weekly affective labels aggregate the repeated self-reports used in the momentary track, while the corresponding features are

rolled up from daily summaries. The weekly track therefore changes both the prediction unit and the feature-history window. A weekly result should not be interpreted as a longer-horizon forecast, but as prediction of the same underlying measurements at a coarser temporal resolution. K-EmoPhone provides the clearest example: its study lasts approximately one week, and only 13.8% of its labels have a complete seven-day history. Its weekly cells are retained and reported with this limitation rather than excluded.

A.7 COVERAGE

Coverage is ragged by design. A resource can appear under S3 without a within-study cell at the same timescale, because at S3 it is the study held out rather than the study modelled — StudentLife's weekly stress cells exist only in that form. CES and GLOBEM never appear under S3 at all: CES because a CES row is a different unit, and GLOBEM because its two targets, PHQ-4 and the depression score, have no transfer partner in the benchmark, the one nominal pairing, PHQ-4 with CES, being deliberately not run (Appendix A.4). Screen-on transfer is run over the directions of the valence roster.

Table 11: **Coverage: dataset × target × timescale.** Digits name the settings that produced reportable cells — 1 within-user temporal, 2 within-dataset unseen users, 3 cross-dataset; a dash marks a combination the benchmark does not evaluate. The screen-on rows give the behavioral target's roster, which differs from the affective rosters at every timescale; DeepStress carries no screen stream. StudentLife's weekly stress enters only as an S3 target, because its own weekly cells fall below the reporting floor (Table 12). Screen-on has no S3 cell for StudentLife, whose release keeps only episodes of at least an hour and so measures a different quantity, nor for CES and GLOBEM, which share too few feature columns with the other studies (Appendix C.3). The screen-on construction, including the S1 training-slice median, is in Appendix A.4.

Resource	Target	Momentary	Daily	Weekly
K-EmoPhone	stress	1 2 3	1 2 3	1 2 3
EmoPhone	stress	1 2 3	1 2 3	1 2 3
DeepStress	stress	1 2 3	1 2 3	1 2 3
StudentLife	stress	1 2 3	1 2 3	3
CES	stress	–	1 2	1 2
K-EmoPhone	valence	1 2 3	1 2 3	1 2 3
EmoPhone	valence	1 2 3	1 2 3	1 2 3
DiversityOne	valence	1 2 3	1 2 3	1 2 3
K-EmoPhone	arousal	1 2 3	1 2 3	1 2 3
EmoPhone	arousal	1 2 3	1 2 3	1 2 3
GLOBEM	phq4	–	–	1 2
GLOBEM	dep	–	–	1 2
K-EmoPhone	screen-on	1 2 3	1 2 3	–
EmoPhone	screen-on	1 2 3	1 2 3	1 2 3
StudentLife	screen-on	1 2	1 2	1 2
CES	screen-on	–	1 2	1 2
DiversityOne	screen-on	1 2 3	1 2 3	1 2 3
GLOBEM	screen-on	–	–	1 2
K-EmoPhone	3 other items	1 2	–	–
EmoPhone	14 other items	1 2	–	–

A.8 FINAL MODELLING SAMPLE SIZES

Settings 1 and 2 are given in full below; setting 3 holds out a whole study and its cells are keyed by transfer direction rather than by dataset, so its samples are reported with the directions in Appendix C. Every target is binary and every row is therefore a class count: the benchmark fits no regression. A reporting floor and an advisory line apply: a cell is reported only with both classes present and at least 30 test rows, and a cell whose any single fold falls below 200 test rows is marked underpowered and read as indicative.

Table 12: **Complete benchmark sample counts.** Every reportable within-resource cell, with S1 shown on the left and S2 on the right. S3 holds out an entire resource, so its cells are indexed by transfer direction and reported in Appendix C. *Prev.* denotes the proportion of test observations in the positive class. A cell is reportable only when both classes are present and the test set contains at least 30 observations; under S2 the floor applies to each fold's test set. Dashes indicate cells excluded by this reporting rule: StudentLife's weekly stress has 7 test rows under S1 and fewer than 30 in every S2 fold, so it reports only as an S3 transfer partner. The screen-on rows closing each timescale block are the sensed target of Appendix A.4; its label comes from the device rather than a response, which is why StudentLife's weekly screen-on cells clear the rule that its weekly stress does not.

Resource	Target	S1				S2			
		Users	Train	Test	Prev.	Users	Train	Test	Prev.
Momentary									
K-EmoPhone	stress	77	3,199	1,106	0.54	77	3,402	5,381	0.53
K-EmoPhone	valence	71	2,982	1,031	0.81	77	3,436	5,381	0.79
K-EmoPhone	arousal	74	3,122	1,079	0.53	77	3,412	5,381	0.56
EmoPhone	stress	294	29,911	10,126	0.44	297	33,140	51,941	0.46
EmoPhone	valence	293	29,859	10,108	0.79	297	33,558	51,941	0.77
EmoPhone	arousal	296	30,129	10,200	0.46	297	33,333	51,941	0.47
DeepStress	stress	24	3,322	1,122	0.37	24	4,563	7,658	0.31
StudentLife	stress	32	765	268	0.72	35	1,170	1,795	0.76
DiversityOne	valence	219	65,529	21,961	0.96	348	102,488	160,607	0.97
K-EmoPhone	screen-on	75	3,130	1,081	0.52	76	3,400	5,305	0.50
EmoPhone	screen-on	291	29,606	10,023	0.52	292	32,138	51,116	0.50
StudentLife	screen-on	33	2,535	863	0.47	35	4,112	6,772	0.44
DiversityOne	screen-on	347	96,171	32,243	0.47	347	103,862	160,546	0.51
Daily									
K-EmoPhone	stress	38	130	70	0.54	76	303	481	0.54
K-EmoPhone	valence	16	58	30	0.73	76	310	494	0.90
K-EmoPhone	arousal	41	143	76	0.43	77	303	480	0.59
EmoPhone	stress	241	3,483	1,290	0.44	297	4,574	7,175	0.45
EmoPhone	valence	188	2,834	1,052	0.78	297	4,694	7,341	0.84
EmoPhone	arousal	237	3,413	1,274	0.41	297	4,621	7,129	0.45
DeepStress	stress	20	471	169	0.35	24	600	954	0.29
StudentLife	stress	36	494	184	0.79	47	674	1,075	0.78
CES	stress	205	19,411	6,571	0.55	205	20,539	32,479	0.49
DiversityOne	valence	24	300	119	0.94	331	3,655	5,769	0.99
K-EmoPhone	screen-on	72	244	139	0.37	75	296	463	0.46
EmoPhone	screen-on	297	4,433	1,638	0.44	297	4,840	7,593	0.49
StudentLife	screen-on	46	1,008	356	0.58	46	1,066	1,704	0.49
CES	screen-on	203	19,311	6,540	0.60	203	20,505	32,313	0.50
DiversityOne	screen-on	292	3,188	1,159	0.45	317	3,498	5,490	0.48
Weekly									
K-EmoPhone	stress	59	716	268	0.49	71	901	1,416	0.56
K-EmoPhone	valence	49	600	227	0.77	71	911	1,416	0.79
K-EmoPhone	arousal	61	734	276	0.53	71	885	1,416	0.55
EmoPhone	stress	285	23,683	8,050	0.44	297	25,946	40,575	0.46
EmoPhone	valence	280	23,580	8,012	0.79	297	26,115	40,575	0.77
EmoPhone	arousal	288	23,704	8,057	0.45	297	26,018	40,575	0.46
DeepStress	stress	24	6,648	2,231	0.34	24	7,107	11,096	0.36
StudentLife	stress	–	–	–	–	–	–	–	–
CES	stress	212	19,507	6,613	0.55	213	20,964	32,654	0.48
DiversityOne	valence	171	45,832	15,368	0.96	295	74,527	117,935	0.97
GLOBEM	phq4	151	808	300	0.46	549	3,217	5,046	0.17
GLOBEM	dep	340	1,892	818	0.52	703	4,401	6,887	0.47
EmoPhone	screen-on	271	21,095	7,173	0.28	295	23,128	35,859	0.47
StudentLife	screen-on	8	119	41	0.41	11	160	237	0.22
CES	screen-on	211	19,105	6,481	0.65	211	20,549	31,991	0.50
DiversityOne	screen-on	251	59,005	19,802	0.33	272	63,349	99,464	0.44
GLOBEM	screen-on	692	3,678	1,570	0.33	700	4,202	6,583	0.47

A.9 DATASET-SPECIFIC CONVERSION

Five of the resources enter as raw event logs processed through RAPIDS under one configuration, and GLOBEM and CES at feature level, shipping pre-extracted aggregates. That difference has a cost the shared spaces inherit: extracted and rolled feature provenance are not interchangeable, and the benchmark keeps them from sharing a feature space where the difference would be confounded with the comparison.

Table 13: **Dataset-specific raw-to-feature conversion.** Five resources enter as raw event logs processed through RAPIDS under one configuration; GLOBEM and CES enter at feature level. Each departure from the default data-yield rule has a measured reason.

Resource	How features are produced	Row-quality rule	Why that rule
K-EmoPhone	raw event logs extracted by RAPIDS into event windows	yield	default: data yield ≥ 0.33
EmoPhone	raw event logs extracted by RAPIDS, one run per wave	yield	default: data yield ≥ 0.33
DeepStress	raw event logs extracted by RAPIDS into event windows	yield	default: data yield ≥ 0.33
StudentLife	raw event logs extracted by RAPIDS into event windows	sensor_presence	periodic sampling makes yield structurally low; the default rule would drop about 93% of its rows
CES	release's precomputed hourly and daily tables; no raw extraction	screen_presence	carries none of the presence rule's columns, so that rule matched nothing and kept every row; screen is its one strong family
DiversityOne	converted to an AWARE-shaped layout first, per collection site and time zone, then extracted by RAPIDS	yield	default: data yield ≥ 0.33
GLOBEM	none: the release ships day-level features already computed	none	no rule defined; the release ships no data-yield column

A.10 FEATURE CONSTRUCTION BY TIMESCALE

The same canonical features are computed over three different windows, and the window is part of the feature's identity rather than a parameter of it: a momentary feature summarizes the 60 minutes before a prompt, a daily feature summarizes a local calendar day, and a weekly feature summarizes a 7- or 14-day history which, on resources that ship no extracted weekly table, is rolled up from that window's daily features. Temporal alignment between feature and target windows is timescale- and dataset-specific and is documented in Table 10.

A.11 NATIVE FEATURE SPACES

A Native space contains every benchmark-valid feature available within one resource, and it is not one number per resource: it changes with the timescale, because the daily and weekly tables summarize different windows and drop columns that carry nothing over them. The counts range from 48 columns for CES at the daily window to 201 for GLOBEM at the weekly one. A native count can exceed the 166-feature canonical schema of Table 7, because a weekly table repeats one canonical feature over its 7-day and 14-day history windows: GLOBEM's 201 columns are that repetition rather than 201 distinct behavioural quantities.

Table 14: **Native feature counts, by resource and timescale.** The feature count a model received within one resource; a dash marks a track the resource does not enter. The ordering is not a ranking of sensing richness — GLOBEM's 201 columns are one vocabulary repeated over history windows.

Resource	Momentary	Daily	Weekly
K-EmoPhone	85	86	51
EmoPhone	114	114	144
DeepStress	53	53	58
StudentLife	120	124	58
CES	–	48	97
DiversityOne	88	88	108
GLOBEM	–	–	201

A.12 SHARED CROSS-DATASET FEATURE SPACES

A Shared space begins from the intersection of semantically compatible features available in the resources participating in a target–timescale roster. This first step is structural: feature availability is determined from the declared feature schemas and conversion mappings and does not inspect feature values from held-out observations. A feature absent from any participating resource is therefore excluded from the corresponding Shared schema.

Value-dependent validity checks are applied only after the evaluation split is defined. Features exceeding the benchmark missingness threshold or having zero variance on the training side are removed before model fitting. Under S1 and S2, these checks use the corresponding training split. Under S3, they use source training data only; feature values from the held-out target dataset are not used to determine feature eligibility. Consequently, source-only supervised and domain-generalization methods do not use unlabeled target feature values during feature construction or preprocessing. Unsupervised domain-adaptation methods receive target features only subsequently, as part of the adaptation procedure described in Appendix B.1.

The Shared schema is defined per (timescale, target) roster, while the final set of model inputs may be reduced by training-side validity checks for a particular evaluation cell.

Table 15: **Shared cross-dataset schemas and evaluation coverage.** One row per target–timescale Shared schema used by the benchmark. *Schema* is the number of semantically compatible features available across all resources in the corresponding roster, determined from feature definitions and availability only. It therefore does not inspect feature values from held-out observations. After an evaluation split is defined, value-dependent missingness and zero-variance filtering is fitted on the training side of each evaluation cell (Table 16); consequently, the final number of model inputs may be smaller than the structural schema size and may differ across folds or transfer directions. *S1* and *S2* count resources with a reportable cell (weekly StudentLife is a transfer partner only: its own weekly cells fall below the reporting floor); *S3* counts transfer directions, pairwise + pooled. The screen-on rows correspond to the S3 spaces of the sensed target in Appendix A.4; its S1 and S2 cells use the Native representation only. StudentLife, CES, and GLOBEM are outside its S3 roster for the reasons described in Appendix C.3. *Runs* sums the corresponding benchmark cells over S1 (14 models; domain-generalization and domain-adaptation methods do not apply within a participant), S2, and S3 (27 models each). Across the affective rows there are 3,253 runs in total, of which S3 contributes 27 per transfer direction and target, or 729 per timescale (Table 18); the screen-on rows add 243, 243, and 54 S3 runs. A *run* is one fitted model result for one (setting, timescale, dataset or transfer direction, target, feature space, model), aggregated over folds and seeds. Only the 27 benchmark models are counted; the majority-class baseline is excluded, and a cell counts only when it is reportable.

Timescale	Target	Participating resources	Schema	S1	S2	S3	Runs
Momentary	stress	DeepStress, EmoPhone, K-EmoPhone, StudentLife	69	4	4	12 + 4	596
Momentary	valence	DiversityOne, EmoPhone, K-EmoPhone	85	3	3	6 + 3	366
Momentary	arousal	EmoPhone, K-EmoPhone	95	2	2	2	136
Momentary	screen-on	DiversityOne, EmoPhone, K-EmoPhone	85	–	–	6 + 3	243
Daily	stress	DeepStress, EmoPhone, K-EmoPhone, StudentLife	69	4	4	12 + 4	596
Daily	valence	DiversityOne, EmoPhone, K-EmoPhone	85	3	3	6 + 3	366
Daily	arousal	EmoPhone, K-EmoPhone	95	2	2	2	136

Timescale	Target	Participating resources	Schema	S1	S2	S3	Runs
Daily	screen-on	DiversityOne, EmoPhone, K-EmoPhone	85	–	–	6 + 3	243
Weekly	stress	DeepStress, EmoPhone, K-EmoPhone, StudentLife	25	3	3	12 + 4	555
Weekly	valence	DiversityOne, EmoPhone, K-EmoPhone	46	3	3	6 + 3	366
Weekly	arousal	EmoPhone, K-EmoPhone	52	2	2	2	136
Weekly	screen-on	DiversityOne, EmoPhone	100	–	–	2	54

A.13 MISSINGNESS AND PREPROCESSING

Preprocessing is a fixed sequence applied identically to every model in a cell: a sensor-coverage filter on the row, a participant eligibility filter, zero-filling in raw units for event-like variables whose absence has a meaningful zero, feature-validity filtering, standardization, clipping at a high quantile of the standardized value, and median imputation with a missingness indicator per sensor family. Structural feature alignment is determined from the declared feature schemas and does not inspect held-out feature values. All value-dependent filtering and preprocessing statistics are estimated from the training side only. In S3, source-only supervised and domain-generalization methods therefore do not use held-out target feature values during preprocessing; unsupervised domain-adaptation methods observe unlabeled target features only as part of their adaptation procedure.

Table 16: **Missingness and the preprocessing pipeline.** Upper: per-resource missingness in the built tables. Lower: shared-feature alignment, value-dependent feature filtering, and the preprocessing sequence applied identically to every model within an evaluation cell. Structural feature alignment uses only feature definitions and availability and does not inspect held-out feature values. All value-dependent feature filtering and preprocessing statistics are estimated from the training side only. In S1, standardization is per-participant and uses only each participant's training slice. In S2 and S3, held-out participants have no training slice, so they are standardized with pooled statistics estimated from the training participants. Under S3, source-only supervised and domain-generalization methods do not use held-out target feature values for feature filtering or preprocessing.

Per-resource missingness

Resource	Feature columns	Mean missing	All-empty columns	Median data yield
K-EmoPhone	101	34%	12	1.00
EmoPhone	133	51%	0	1.00
DeepStress	69	34%	12	0.40
StudentLife	151	81%	11	0.05
CES	48	71%	0	0.07
DiversityOne	97	46%	7	1.00
GLOBEM	204	20%	0	–

Feature-space filtering and preprocessing

Stage	What	How	Fitted on
Schema	Shared-feature alignment	Define the candidate Shared schema as the intersection of semantically compatible features available across the datasets in the corresponding target–timescale roster. This step uses feature definitions and availability only and does not inspect feature values.	No fitted statistics.
0	Shared-space validity filter	After structural alignment, remove features that fail the benchmark's 99%-missing criterion or have zero variance on the training side of the evaluation cell. Under S3, these checks use source training data only and do not inspect held-out target feature values.	Training side of the evaluation cell; source training data only under S3.
1	Row-quality filter	Apply the resource-specific row-quality rule; by default, retain rows with data yield ≥ 0.33.	Nothing fitted.
2	Minimum responses per participant	Remove participants with fewer than 20 labelled observations.	Nothing fitted.

Feature-space filtering and preprocessing continued

Stage	What	How	Fitted on
3	Zero-fill structural absence	Replace structurally absent event-count features with zero in their raw units before standardization, because the absence of an event represents a measured zero.	Nothing fitted.
4	Standardization (z-score)	For S1, each participant is standardized using statistics estimated from that participant's training slice. For S2 and S3, held-out participants have no participant-specific training slice, so they are standardized using the pooled mean and standard deviation estimated from the training participants.	S1: each participant's training slice. S2 and S3: training split, pooled over training participants.
5	Outlier clipping	Clip standardized values at the 0.999 quantile of $\|z\|$, with a minimum clipping threshold of 10.	Training split.
6	Median imputation and missingness indicators	After standardization and clipping, replace remaining missing values using training-set medians and add a missingness indicator for each feature family.	Training split.

B BENCHMARK PROTOCOLS, MODELS, AND COMPUTE

B.1 EVALUATION PROTOCOLS

Section 4.1 defines the three evaluation settings and Table 17 states them formally. This appendix describes the procedures that implement them. Every fraction, fold count, seed, and reporting floor reported below was taken from the code that produced the results rather than from the protocol description, so the pseudocode matches the runs exactly.

The three settings differ in what they withhold and in what they make available, not only in how difficult they are. Four differences matter for interpreting the results and are easily lost in summary: which slice validation is drawn from, whether target labels exist at all, whether unlabelled target features are available, and what the domain-generalization objective treats as a domain.

The last point is the one most often misread, so we state it directly. Two distinct things are called domains here. The *held-out evaluation domain* is what a setting withholds from training: whole participants under S2, and one whole dataset under S3. The *domain labels supplied to the objective* are the grouping variable the loss reads during training. In our experiments this grouping is the source participant under S3, exactly as under S2. S3 therefore measures transfer to a held-out dataset while its domain-generalization methods partition the source pool by participant. Source datasets are never used as domain labels.

Table 17: **The evaluation protocol, by setting.** What each setting holds out, trains on, validates on and reports, together with the access rules, preprocessing, tuning budget and training loop that apply to it. A row spanning all three columns is identical in every setting. An entry marked *n/a* is a row the setting does not define, not one left unfilled: under S1 training and test rows come from the same participants, so no shift exists for a domain-generalization or adaptation objective to close.

Property	S1 within-user	S2 unseen users	S3 unseen study
SPLIT GEOMETRY			
What is held out	later rows of a known participant	whole participants of one study	one whole study
Training rows	that participant's earliest 60%	the fold's training participants	one source study, or all others pooled
Validation	the participant's middle 20%	grouped 20% holdout of the fold's training participants	grouped 20% holdout over source participants
Test rows	the participant's latest 20%	the fold's held-out participants	every row of the held-out study
Folds	one split	5, grouped by participant, stratified by label	one split
Split seed	42, fixed, so every model in a cell sees identical splits		
S1 history cap	30 days at momentary only; uncapped at daily and weekly	n/a	n/a
ELIGIBILITY			
Participant dropped when	any slice is empty, or the training slice holds one class	n/a — participants are held out whole	n/a — studies are held out whole
Responses per participant	≥ 20 at momentary and daily; no gate at weekly		
ACCESS TO THE HELD-OUT SIDE			
Target labels	n/a — same participants	n/a — same study	never: not for fitting, tuning, checkpointing or the threshold

Property	S1 within-user	S2 unseen users	S3 unseen study
Unlabelled target features	n/a	n/a	source-only and DG: no. UDA: yes, the test-split feature matrix
Domain labels read by the objective	n/a — no shift to close	source participant	source participant, never the source dataset
Standardisation	that participant's training slice	training participants, pooled	source training split, pooled
Feature space	Native and Shared	Native and Shared	Shared only
MODELS AND TUNING			
Families defined	baselines, deep tabular, pretrained tabular (14 models)	all five (27 models)	all five (27 models)
Tuning budget	30 Optuna trials per (setting, target, dataset-or-direction, model), tree-structured Parzen, fixed sampler seed, selected on validation AUROC		
Where the search runs	on the single split	independently within each outer fold; selected hyperparameters reused across seeds within that fold	strictly source-only
Pretrained tabular models	untuned by design; the budget is left unspent and the cell records it		
TRAINING LOOP			
Optimiser	AdamW, cross-entropy without reduction		
Batch size	512 for training and evaluation alike		
Epochs	at most 50; early stopping, patience 20, minimum improvement 10^{-4}, best epoch restored		
WHAT IS REPORTED			
Seeds	42, 43, 44 — three per cell; no cell was re-executed, and a seed is not a repeat		
Fits per cell	3	15, three seeds on each of five folds	3
Dispersion reported	SD across 3 seeds	SD across 5 folds of seed-averaged values	SD across 3 seeds
Reporting floor	$\geq$ 30 test rows and both classes present; any fold below 200 test rows is flagged underpowered		
Deployment scenario it stands for	a known user's future	a new user of a known study	a new study

B.2 THE THREE SPLIT PROCEDURES

Each setting is implemented by one split procedure. The three share a preprocessing pipeline and a training loop, and differ only in what they hold out and in what the held-out side makes available.

S1 — within-user temporal prediction. Each participant's observations are ordered in time and cut at 60% and 80% of their length: the earliest 60% trains, the next 20% validates, and the latest 20% tests. The division is positional rather than calendar-based, so a participant who answered irregularly is split in the same proportions as one who answered steadily.

Algorithm 1 S1 — within-user temporal split

Require: rows of one dataset–target pair with `pid`, `window_start`, label y; day cap D (30 at momentary, else ∞); gate m (20 at momentary and daily, else 0)
Ensure: train/validation/test slices and the dropped participants X

for each participant p **do**
  $R_p \leftarrow$ rows of p sorted by `window_start`, truncated to the first D days
  $n \leftarrow |R_p|$; $(\text{tr}_p, \text{va}_p, \text{te}_p) \leftarrow (R_p[0{:}\lfloor .6n \rfloor],\ R_p[\lfloor .6n \rfloor{:}\lfloor .8n \rfloor],\ R_p[\lfloor .8n \rfloor{:}])$ ▷ positional; y never read
  drop unlabelled rows from each slice
  if p has fewer than m labelled rows, any slice is empty, or tr_p is single-class **then**
    $X \leftarrow X \cup \{p\}$ with its reason ▷ no per-participant score exists
  end if
end for
$(\text{Tr}, \text{Va}, \text{Te}) \leftarrow$ concatenation of the surviving slices; **abort** if none survive
mark the cell NOTREPORTED if $|\text{Te}| < 30$ or single-class, UNDERPOWERED if $|\text{Te}| < 200$ ▷ Table 17
return $(\text{Tr}, \text{Va}, \text{Te})$, X

Algorithm 1. S1, within-user temporal prediction. Splitting one dataset–target pair into train, validation and test slices, participant by participant, with the eligibility gates and reporting floors that apply to the cell.

Two properties of this construction are deliberate. First, the cut is computed from timestamps alone and never from the label. A split geometry that depended on the quantity being predicted would be circular wherever a label definition is itself fitted on the training slice. Second, a participant is dropped when any of their three slices is empty or when their training slice holds a single class, since no per-participant score exists in either case. The dropped identifiers are recorded rather than discarded, and the retained counts appear in Appendix A.2.

One parameter varies by timescale. On the momentary track, each participant's history is capped at their first 30 days, following (Zhang et al., 2026b). The daily and weekly tracks are uncapped: a person-day or person-week series is short enough that the cap would either bind on nobody or bind severely. On one capped weekly run it retained 215 of 703 participants while still reporting the affected cells as usable. A minimum of 20 labelled responses per participant applies at the momentary and daily tracks. No participant gate applies at the weekly track, where a row counts person-weeks rather than prompts.

Because the same participants appear in all three slices, only the time index separates them. S1 therefore measures temporal generalization under user-informed deployment and carries no information about generalization to unseen users.

S2 — unseen users within a dataset. Participants are partitioned into five folds, grouped on participant identity and stratified on the label, so that class balance is preserved across folds while no participant is divided. The fold assignment is fixed, so every model in a cell sees identical folds and a model-to-model difference is never a fold-assignment difference.

Algorithm 2 S2 — unseen users, same study

Require: rows of one dataset–target pair with `pid` and label y
Ensure: five participant-disjoint (train, validation, test) splits
$F \leftarrow$ StratifiedGroupKFold$(5,\ \text{shuffle},\ \text{seed}{=}42)$, grouped by `pid`, stratified by y
for each fold $(\text{tr}, \text{Te}) \in F$ **do**
  $(\text{Tr}, \text{Va}) \leftarrow$ GroupedHoldout$(\text{tr}, 0.2)$ ▷ tuning never sees a test participant
  emit $(\text{Tr}, \text{Va}, \text{Te})$
end for

function GroupedHoldout(I, f)
  $U \leftarrow$ participants appearing in I
  if $|U| < 2$ **then return** a stratified split of the *rows* of I
  end if
  label each $p \in U$ by its own majority class
  hold out a fraction f of *participants*, stratified on that label where possible
  return the rows of the kept and of the held-out participants
end function

Algorithm 2. S2, unseen users within a dataset. Five-fold cross-validation grouped by participant and stratified by the label, with validation carved from each fold's training participants by the same grouped rule. Participants rather than rows are held out at every level.

Validation is a grouped 20% holdout carved from that fold's training participants and stratified by each participant's own majority label, which leaves the validation participants disjoint from both the training and the test participants of that fold. Participants rather than rows are held out at every level, and this is what makes the setting a test of transfer to an unseen person rather than of interpolation within one.

S3 — unseen study. The target dataset is held out in its entirety and becomes the test set. Training draws either on one other dataset, which we call a *pairwise* direction, or on every other dataset in that target's roster, a *pooled* direction, which is the leave-one-dataset-out case. In pooled directions the source datasets enter with inverse-size sample weights, so that a single large study cannot dominate a training set assembled from several.

Algorithm 3 S3 — unseen study

Require: pooled rows of a dataset group over the shared feature space; target dataset T; source subset S (all datasets but T when omitted)
Ensure: one (train, validation, test) split for the direction $S \to T$
Te $\leftarrow$ every row whose dataset is T; Src $\leftarrow$ rows of S ▷ **abort** if either is empty
$(\text{Tr}, \text{Va}) \leftarrow$ GroupedHoldout$(\text{Src}, 0.2)$ ▷ source participants only
if $|S| > 1$ **then**
  weight each $s \in S$ by $1/|\text{rows}(s)|$ ▷ no single study dominates the pool
end if
$d \leftarrow$ `pid` of each row of Tr ▷ domain labels for DG: the source *participant*, never s
give the adaptation methods the features of Te, unlabelled ▷ transductive; labels of T are never read
return $(\text{Tr}, \text{Va}, \text{Te})$, d, labelled $S \to T$

Algorithm 3. S3, unseen study. One dataset is held out entire and becomes the test set; training draws on one other dataset (pairwise) or on every other dataset of the target's roster (pooled, the leave-one-dataset-out case). The structural Shared schema is defined once per target and timescale, while value-dependent validity filtering is fitted separately for each direction using source training data only.

The enumeration is all-but-one rather than the full sweep of source subsets, which would multiply the direction count without answering a different question. At two datasets the pooled form degenerates into the pairwise one, and the duplicate is dropped. Two counts follow and are easily conflated: 27 (direction, target) triples are run per timescale, and these reduce to 23 distinct directions once the dataset pairs shared between targets are deduplicated. The EmoPhone–K-EmoPhone pair, for

instance, is a direction of the stress, valence and arousal rosters alike, so its two directions serve three rosters. The direction roster is identical at all three timescales.

Table 18: **The 23 S3 transfer directions, as run.** Each row is one transfer direction: the study held out as the test set, the source studies trained on, and the prediction targets that direction serves. Pairwise directions train on one source study, pooled directions on every other study of that target's roster. The affective roster is identical at all three timescales. Screen-on, the sensed target of Appendix A.4, has the valence roster (EmoPhone, K-EmoPhone and DiversityOne), so it adds no direction of its own: it serves nine of the valence directions at the momentary and daily timescales and, because K-EmoPhone has no weekly screen-on label, only the two between EmoPhone and DiversityOne at the weekly one (‡ marks a direction whose screen-on runs stop short of weekly). *Runs* counts the runs a direction contributes at each timescale (*M*omentary, *D*aily, *W*eekly), 27 per target served. The affective runs are 729 per timescale and 2,187 over the three, the S3 share of Table 15's runs, as defined there; screen-on adds 243, 243 and 54, for 972, 972 and 783 in all.

			Runs		
Source	Held-out target dataset	Prediction targets	M	D	W
Pairwise, one source study (16)					
EmoPhone	DeepStress	stress	27	27	27
K-EmoPhone	DeepStress	stress	27	27	27
StudentLife	DeepStress	stress	27	27	27
EmoPhone	DiversityOne	valence, screen-on	54	54	54
K-EmoPhone	DiversityOne	valence, screen-on‡	54	54	27
DeepStress	EmoPhone	stress	27	27	27
DiversityOne	EmoPhone	valence, screen-on	54	54	54
K-EmoPhone	EmoPhone	arousal, stress, valence, screen-on‡	108	108	81
StudentLife	EmoPhone	stress	27	27	27
DeepStress	K-EmoPhone	stress	27	27	27
DiversityOne	K-EmoPhone	valence, screen-on‡	54	54	27
EmoPhone	K-EmoPhone	arousal, stress, valence, screen-on‡	108	108	81
StudentLife	K-EmoPhone	stress	27	27	27
DeepStress	StudentLife	stress	27	27	27
EmoPhone	StudentLife	stress	27	27	27
K-EmoPhone	StudentLife	stress	27	27	27
Pooled, every other study of the roster (7)					
Sources	Held-out target dataset	Prediction targets	M	D	W
EmoPhone, K-EmoPhone, StudentLife	DeepStress	stress	27	27	27
EmoPhone, K-EmoPhone	DiversityOne	valence, screen-on‡	54	54	27
DeepStress, K-EmoPhone, StudentLife	EmoPhone	stress	27	27	27
DiversityOne, K-EmoPhone	EmoPhone	valence, screen-on‡	54	54	27
DeepStress, EmoPhone, StudentLife	K-EmoPhone	stress	27	27	27
DiversityOne, EmoPhone	K-EmoPhone	valence, screen-on‡	54	54	27
DeepStress, EmoPhone, K-EmoPhone	StudentLife	stress	27	27	27

Target labels are unavailable at every stage: fitting, hyperparameter selection, checkpoint selection, and the decision threshold used by the threshold-dependent secondary metrics. Preprocessing statistics, including standardization, are fitted on the training split alone (Appendix A.13), so source-only supervised and domain-generalization models never observe held-out participants' features at any stage: not for preprocessing, model fitting, model selection, or adaptation. Domain-adaptation methods additionally receive the target's test-split feature matrix without labels as input to the adaptation procedure; their setting is therefore transductive, and we state this wherever their results appear. S3 additionally requires the shared feature space, since a transfer direction has no meaning over

columns one study does not collect, so its performance reflects both distribution shift and the constraint of cross-study feature alignment. The shift it exposes is compound: participants, protocol, study environment, collection era, feature distributions, and potentially label prevalence all change together. It should not be read as a test of geography, culture, or device in particular.

B.3 Seeds, folds, repeats and aggregation

Three quantities are kept distinct, since conflating any two would misstate the uncertainty. A *seed* varies model initialisation and, for the in-context models, the ensemble's preprocessing draws; three seeds are run per cell. A *fold* is a participant-disjoint partition and exists only under S2, which uses five. A *repeat* is an independent re-execution of a cell, and none was performed. A seed is not a repeat, and the two are never combined in an uncertainty estimate. A cell therefore rests on three fits under S1 and S3, and on fifteen under S2.

Aggregation follows a fixed order: each seed is scored on its split, seeds are averaged, folds are averaged, and the result is the cell. Dataset–target pairs are never averaged into a single number, because they are distinct prediction problems. The reported dispersion is therefore not the same quantity in every setting. It is the standard deviation across seeds under S1 and S3, and the standard deviation across folds of already seed-averaged values under S2, and it is labelled accordingly wherever it appears. Because the split is fixed before the seed loop, all seeds share one test split, so a reported test size is that split's size rather than a sum across seeds. A fold that fails invalidates its entire cell. We never report a mean over the surviving folds, since that would answer a different question from the one the cell poses.

B.4 The model roster

The benchmark evaluates 27 models in five families: conventional baselines, deep tabular architectures, pretrained tabular models, domain-generalization methods and unsupervised domain-adaptation methods. All 27 are evaluated under S2 and S3. Fourteen are evaluated under S1 — the baselines, the deep tabular models and the pretrained models — because domain generalization and domain adaptation are undefined where training and test rows come from the same participants. There is no distribution shift for either family to close.

Table 19: **The 27-model roster and its per-cell coverage.** Which models report in which setting and timescale, over the core affective and mental-health targets. • reported, ○ undefined by construction. †TabICL and TabDPT were not run on the momentary extended targets, so the extended blocks of Tables 29 and 31 carry twelve supervised models rather than fourteen, and these two models have 204 fewer fits than the other pretrained models (Table 27).

		Momentary			Daily			Weekly		
Model	Family	S1	S2	S3	S1	S2	S3	S1	S2	S3
RF	Baseline	•	•	•	•	•	•	•	•	•
XGBoost		•	•	•	•	•	•	•	•	•
LightGBM		•	•	•	•	•	•	•	•	•
MLP		•	•	•	•	•	•	•	•	•
ResNet		•	•	•	•	•	•	•	•	•
TabNet	Deep Tabular	•	•	•	•	•	•	•	•	•
DCN		•	•	•	•	•	•	•	•	•
SAINT		•	•	•	•	•	•	•	•	•
TabTransformer		•	•	•	•	•	•	•	•	•
FT-Transformer		•	•	•	•	•	•	•	•	•
TabPFN	Tabular Foundation	•	•	•	•	•	•	•	•	•
TabFM-Torch		•	•	•	•	•	•	•	•	•
TabICL†		•	•	•	•	•	•	•	•	•
TabDPT†		•	•	•	•	•	•	•	•	•
ERM	Domain Generalization	○	•	•	○	•	•	○	•	•

Model	Family	Momentary S1	Momentary S2	Momentary S3	Daily S1	Daily S2	Daily S3	Weekly S1	Weekly S2	Weekly S3
IRM		○	●	●	○	●	●	○	●	●
VREx		○	●	●	○	●	●	○	●	●
GroupDRO		○	●	●	○	●	●	○	●	●
CSD		○	●	●	○	●	●	○	●	●
MixStyle		○	●	●	○	●	●	○	●	●
Fish		○	●	●	○	●	●	○	●	●
DANN	Unsupervised Domain Adaptation	○	●	●	○	●	●	○	●	●
CDAN		○	●	●	○	●	●	○	●	●
DAN		○	●	●	○	●	●	○	●	●
JAN		○	●	●	○	●	●	○	●	●
CORAL		○	●	●	○	●	●	○	●	●
MCC		○	●	●	○	●	●	○	●	●

Three properties are held constant, so that a difference between families is attributable to the method rather than to its treatment. Every model in a cell receives the same rows, feature matrix, label vector, preprocessing transform and split. Every tunable model receives the same budget of 30 Optuna trials, selected on validation AUROC. Every model trained by backpropagation runs one shared loop, and the domain-generalization and domain-adaptation methods additionally share the baseline multilayer perceptron's encoder and override only their loss term. This last property is what makes the shift-aware comparisons answerable: optimiser, schedule, batch size, early stopping, seeding and capacity are identical across a method and its control, so a margin reflects the objective rather than the architecture.

Baselines and deep tabular architectures. Random forest, XGBoost, LightGBM, a multilayer perceptron and a residual network form the baseline family, each tuned on the common budget over five to ten hyperparameters. The family establishes whether newer methods add value, and it should not be read as a set of weak controls: tree ensembles are strong methods on structured sensing features, and prior mobile-sensing work has found random forest to outperform several alternatives on mood inference. TabNet, DCN, SAINT, TabTransformer and FT-Transformer represent architectures designed specifically for tabular data. Each is trained from scratch on every benchmark task and carries no external pretraining, which is the axis that separates this family from the pretrained models, and each is tuned on the same budget over six to seven hyperparameters. The question this family answers is whether specialised tabular inductive biases improve on conventional baselines and, where they do, whether the advantage survives a change of evaluation setting.

The pretrained tabular models. Four pretrained models are evaluated. They form their own family because they are the only models carrying external pretraining and because none of the four updates its weights on a benchmark task. All four are frozen in-context learners: the training slice is supplied as context at inference rather than used to update weights. TabFM-Torch caps its in-context set at 10,000 rows, a measured ceiling on the available hardware, whereas TabPFN, TabICL and TabDPT condition on the full training slice. TabDPT differs in its pretraining, which uses real rather than synthetic tables; its optional per-query retrieval is off at the shipped defaults used here (Table 20).

Table 20: **The four pretrained tabular models, as implemented.** Upper: what each model is and how it conditions on the training slice. Lower: the versions the sweep ran. All four are frozen and untuned, fit a binary classification head, and ignore the sample weights that pooled S3 directions supply, which each cell records rather than absorbing silently; no model's feature limit binds on this grid. Only TabFM-Torch caps its context, so on the largest cross-dataset sources it observes a fraction of what the others do. TabFM's pretraining data is undisclosed, as no technical report has been released; the other three state theirs.

Method

Property	TabPFN	TabFM-Torch	TabICL	TabDPT
what it is	transformer pretrained on synthetic tabular priors	in-context learner ensembling over sampled contexts	in-context learner architected for large tables	in-context learner pretrained on real tables; its retrieval option is off at the shipped defaults
conditions on	the whole training table, in context	a capped sample of the training table	the whole training table, in context	the whole training table, in context
pretraining data	synthetic tabular tasks; no benchmark data	undisclosed; no technical report released	synthetic tabular tasks; no benchmark data	real tables (OpenML), self-supervised; no benchmark data
released by	Prior Labs	Google Research; weights under a non-commercial licence	Qu et al.	Layer 6 AI
context limit	none; the training slice is the context	capped, and the cap binds on part of the grid	none; measured unnecessary at this grid's sizes	none; the training slice is the context
preprocessing	benchmark pipeline, then the model's shipped draws	benchmark pipeline; columns passed as a named frame	benchmark pipeline; library offloading at its default	benchmark pipeline, then the model's internal normalizer
ensemble	shipped default	reduced from the shipped 32 to 8, for sweep runtime	shipped default	none; a single forward pass, as the shipped `predict_proba` runs

Environment

Property	TabPFN	TabFM-Torch	TabICL	TabDPT
package version	tabpfn 8.0.8	tabfm 1.0.1	tabicl 2.1.1	tabdpt 1.2.0
torch	2.6.0+cu124	2.4.0+cu121	2.4.0+cu121	2.6.0+cu124
Python	3.11.5	3.11.5	3.11.5	3.11.5

None of the four receives hyperparameter optimisation, which is a protocol decision rather than an omission. These models are advanced on the claim that they require no per-task search, so spending the shared budget on them would evaluate a different method from the one under comparison. The budget is left unspent, and every such cell is recorded as untuned.

Shift-aware objectives. Seven domain-generalization methods are evaluated: an empirical-risk-minimisation control together with IRM, VREx, GroupDRO, CSD, MixStyle and Fish. The objective is robustness to unseen domains without observing target data. Every method is read as a difference from the control rather than ranked against the other methods, and the control is not itself a method.

Participant-level domains carry a practical consequence that belongs with any result from this family. A domain-generalization result depends on how many domains a training batch observes, so we report that count alongside it. A uniformly sampled batch over several hundred participant domains would estimate each per-domain risk from fewer than two rows, which makes the per-domain objectives meaningless and their inner loops intractable. The training loop therefore draws eight domains of 64 rows per step rather than a uniformly sampled batch. Batch size and steps per epoch are unchanged, and the domain count is held equal across every method in the family.

Six unsupervised domain-adaptation methods are evaluated: DANN, CDAN, DAN, JAN, CORAL and MCC. Each adapts using unlabelled target-domain features, and each is compared against a matched source-only model of identical capacity — the baseline multilayer perceptron on the same encoder — rather than against the strongest model in the supervised roster. The access these methods

receive must be stated wherever their results appear. They observe the target's test-split feature matrix without labels, which is the exact set of rows on which they will be scored, so the setting is transductive and a transductive result is not directly comparable to a non-transductive one. Target labels remain unavailable throughout.

B.5 THE TRAINING LOOP AND ITS SEARCH SPACES

Every model trained by backpropagation shares one loop: the neural baselines, the five deep tabular architectures, and all thirteen domain-generalization and domain-adaptation methods. The loop uses AdamW, cross-entropy computed without reduction so that per-sample weights apply exactly, a batch size of 512 for both training and evaluation, at most 50 epochs, and early stopping on validation AUROC with patience 20 at a minimum improvement of 10^{-4}, restoring the best epoch.

The evaluation batch size is fixed rather than left free because SAINT's intersample attention is batch-dependent, so varying it would change that model's predictions. Two further details keep the shift-aware comparison fair. The adversarial adaptation methods place their discriminator in a second optimiser parameter group, so its learning rate is honoured while the encoder keeps the shared loop's optimiser. The domain-generalization methods draw a fixed number of domains per step, as described above.

The 23 tunable models search 150 ranges between them. The neural baselines search learning rate over LogU(10^{-5}, 10^{-2}), weight decay over LogU(10^{-6}, 10^{-3}), hidden width over [64, 512], depth over [1, 8] and dropout over [0, 0.5]. The deep tabular architectures use comparable or narrower ranges.

Table 22: **Baseline search spaces.** The five conventional baselines, with five to ten hyperparameters each, all searched on the shared budget of 30 Optuna trials selected on validation AUROC. Ranges written $\{0\}$ or LogU$(\cdot)$ are searched as a switch between disabling the term and sampling it on a log scale.

Model	Hyperparameter	What it controls	Searched over
RF	n_estimators	trees in the ensemble	[50, 500]
	max_depth	maximum tree depth	[3, 20]
	min_samples_leaf	minimum rows per leaf	[1, 50]
	min_samples_split	minimum rows to split a node	[2, 50]
	max_features	fraction of features considered per split	U(0.3, 1.0)
	class_weight	class re-weighting scheme	{None, balanced, balanced_subsample}
XGBoost	learning_rate	boosting step size	LogU(10^{-5}, 1.0)
	max_depth	maximum tree depth	[3, 10]
	min_child_weight	minimum child weight	LogU(10^{-8}, 10^{5})
	subsample	row subsample per tree	U(0.5, 1.0)
	colsample_bytree	feature subsample per tree	U(0.5, 1.0)
	colsample_bylevel	feature subsample per level	U(0.5, 1.0)
	n_estimators	trees in the ensemble	[50, 500]
	gamma	minimum split loss to divide a leaf	$\{0\}$ or LogU(10^{-8}, 10^{2})
	reg_lambda	L2 penalty	$\{0\}$ or LogU(10^{-8}, 10^{2})
	reg_alpha	L1 penalty	$\{0\}$ or LogU(10^{-8}, 10^{2})
LightGBM	learning_rate	boosting step size	LogU(10^{-3}, 1.0)
	num_leaves	leaves per tree	[8, 128]
	n_estimators	trees in the ensemble	[50, 500]
	min_child_samples	minimum rows per leaf	[5, 100]
	subsample	row subsample per tree	U(0.5, 1.0)
	colsample_bytree	feature subsample per tree	U(0.5, 1.0)
MLP	lr	learning rate	LogU(10^{-5}, 10^{-2})
	weight_decay	AdamW weight decay	LogU(10^{-6}, 10^{-3})
	hidden_dim	encoder width	[64, 512]
	dropout	dropout rate	U(0.0, 0.5)
	num_layers	encoder depth	[1, 8]
ResNet	lr	learning rate	LogU(10^{-5}, 10^{-2})

Model	Hyperparameter	What it controls	Searched over
	weight_decay	AdamW weight decay	LogU(10^{-6}, 10^{-3})
	hidden_dim	encoder width	[64, 512]
	dropout	dropout rate	U(0.0, 0.5)
	num_blocks	residual blocks	[1, 8]

Table 23: **Deep tabular search spaces.** The five architectures designed for tabular data, six to seven hyperparameters each, on the same budget of 30 Optuna trials selected on validation AUROC. The depth and width ranges are narrower than the baselines' because these architectures cost substantially more per fit at equal capacity. None carries external pretraining, which is the axis separating this family from the pretrained tabular models.

Model	Hyperparameter	What it controls	Searched over
TabNet	lr	learning rate	U(10^{-3}, 5×10^{-3})
	weight_decay	AdamW weight decay	LogU(10^{-5}, 5×10^{-4})
	n_da	decision and attention width	{8, 16}
	n_steps	decision steps	[3, 4]
	gamma	feature-reuse coefficient	U(1.1, 1.5)
	lambda_sparse	sparsity penalty	LogU(10^{-5}, 10^{-3})
DCN	lr	learning rate	LogU(5×10^{-5}, 5×10^{-3})
	weight_decay	AdamW weight decay	LogU(10^{-5}, 5×10^{-4})
	num_layers	encoder depth	[1, 2]
	hidden_dim	encoder width	[64, 128]
	num_cross	cross layers	[1, 2]
	dropout	dropout rate	U(0.0, 0.2)
SAINT	d_token	token embedding width	{8, 16}
	lr	learning rate	LogU(2×10^{-4}, 8×10^{-4})
	weight_decay	AdamW weight decay	LogU(10^{-5}, 5×10^{-4})
	n_heads	attention heads	{2, 4}
	num_blocks	transformer blocks	[1, 2]
	attn_dropout	attention dropout	U(0.0, 0.2)
	ffn_dropout	feed-forward dropout	U(0.0, 0.2)
TabTransformer	d_token	token embedding width	{8, 16}
	lr	learning rate	LogU(2×10^{-4}, 8×10^{-4})
	weight_decay	AdamW weight decay	LogU(10^{-5}, 5×10^{-4})
	n_heads	attention heads	{2, 4}
	num_blocks	transformer blocks	[1, 2]
	attn_dropout	attention dropout	U(0.0, 0.2)
	ffn_dropout	feed-forward dropout	U(0.0, 0.2)
FT-Transformer	lr	learning rate	LogU(5×10^{-5}, 5×10^{-4})
	weight_decay	AdamW weight decay	LogU(10^{-5}, 5×10^{-4})
	d_token	token embedding width	{16, 32}
	num_blocks	transformer blocks	[1, 2]
	attn_dropout	attention dropout	U(0.0, 0.2)
	ffn_dropout	feed-forward dropout	U(0.0, 0.2)
	ffn_factor	feed-forward expansion factor	U(1.0, 2.0)

The two shift-aware families draw a shared block plus their own method-specific terms. Domain generalization draws learning rate LogU(3×10^{-5}, 3×10^{-4}), weight decay LogU(10^{-6}, 10^{-2}), dropout from $\{0, 0.1, 0.5\}$, width [64, 512] and depth [1, 8]. Domain adaptation draws an analogous block with learning rate LogU(10^{-5}, 10^{-2}) and weight decay LogU(10^{-6}, 10^{-3}). Because the encoder and the shared block are common to a method and its control, the method-specific terms are the only thing that varies within a family.

Table 24: **Domain-generalization search spaces.** Every method shares the baseline multilayer perceptron's encoder and the shared block below, and overrides only its own loss term, so a margin over the ERM control reflects the objective rather than the architecture. ERM draws the shared block and no method-specific terms, and is tuned on the same budget as the methods it controls for. Domains are source participants in both S2 and S3, and the training loop draws eight of them per step.

Model	Hyperparameter	What it controls	Searched over
Shared block, drawn by every method			
	lr	learning rate	LogU(3×10^{-5}, 3×10^{-4})
	weight_decay	AdamW weight decay	LogU(10^{-6}, 10^{-2})
	dropout	dropout rate	{0.0, 0.1, 0.5}
	hidden_dim	encoder width	[64, 512]
	num_layers	encoder depth	[1, 8]
Method-specific terms			
IRM	penalty	invariance penalty weight	LogU(10^{-1}, 10^{5})
	anneal_exp	penalty annealing epoch, as a power of ten	U(0, 4)
VREx	penalty	variance penalty weight	LogU(10^{-1}, 10^{5})
	anneal_exp	penalty annealing epoch, as a power of ten	U(0, 4)
GroupDRO	eta	group step size	LogU(10^{-3}, 10^{-1})
CSD	csd_lambda	common-specific trade-off	U(0.1, 5.0)
	k	specific-component rank	{2, 3}
MixStyle	alpha	mixing Beta parameter	U(0.1, 0.5)
	p	probability a batch is mixed	U(0.1, 0.9)
	mix	mixing scheme	{random, crossdomain}
Fish	meta_lr	inner-loop meta step size	{0.05, 0.1, 0.5}

Table 25: **Unsupervised domain-adaptation search spaces.** As for domain generalization, every method shares the baseline multilayer perceptron's encoder and the shared block below, overriding only its loss term. Each is compared against a matched source-only model on that same encoder, and each adapts on unlabelled target features while target labels remain unavailable throughout. The adversarial methods place their discriminator in a second optimiser parameter group, so its learning rate is honoured while the encoder keeps the shared loop's optimiser.

Model	Hyperparameter	What it controls	Searched over
Shared block, drawn by every method			
	lr	learning rate	LogU(10^{-5}, 10^{-2})
	weight_decay	AdamW weight decay	LogU(10^{-6}, 10^{-3})
	hidden_dim	encoder width	[64, 512]
	num_layers	encoder depth	[1, 8]
Method-specific terms			
DANN	disc_lr	discriminator learning rate	LogU(10^{-5}, 10^{-2})
	disc_width_exp	discriminator width, as a power of two	[6, 10]
	trade_off	adaptation loss weight	LogU(10^{-2}, 10^{2})
CDAN	disc_lr	discriminator learning rate	LogU(10^{-5}, 10^{-2})
	disc_width_exp	discriminator width, as a power of two	[6, 10]
	trade_off	adaptation loss weight	LogU(10^{-2}, 10^{2})
	entropy	entropy conditioning	{True, False}
	randomized	randomized multilinear map	{False, True}
	randomized_dim	randomized map dimension	{256, 512, 1024}
DAN	trade_off	adaptation loss weight	LogU(0.1, 10.0)
	kernels	Gaussian kernel bandwidths	{3, 5, 7}
	linear	linear-time MMD estimator	{False, True}
JAN	trade_off	adaptation loss weight	U(0.1, 10.0)
	kernels	Gaussian kernel bandwidths	{3, 5, 7}
	linear	linear-time MMD estimator	{False, True}

Model	Hyperparameter	What it controls	Searched over
CORAL	gamma	covariance alignment weight	LogU(0.1, 10.0)
MCC	temperature	class-confusion temperature	U(1.0, 5.0)
	trade_off	adaptation loss weight	LogU(10^{-2}, 10^{1})

B.6 Environment, runtime and compute

Table 26: **Software and hardware environment.** The environment the sweep ran in. The pretrained tabular models each run in their own interpreter because their torch requirements conflict; their versions are in Table 20.

Component	As run
GPU	NVIDIA GeForce RTX 4090, 24 GB, driver 550.163.01
Python	3.11.5
PyTorch	2.4.0+cu121 (CUDA 12.1)
scikit-learn	1.7.2
XGBoost	3.0.5
LightGBM	4.6.0
Optuna	4.6.0
NumPy	1.26.4
pandas	2.3.1
Shared training loop	AdamW, cross-entropy, batch 512, up to 50 epochs, patience 20 on validation AUROC, best epoch restored
Pretrained tabular models	one interpreter each; their versions are in Table 20

The runs behind the reported results come to 13,277 fits and 189 wall-clock hours, a third of which carried a hyperparameter search. The cost is unevenly distributed. The deep tabular family accounts for 70 of those hours against 20 for the baselines, and the pretrained tabular family costs 26 hours despite being untuned, because in-context inference is paid at every fit rather than amortised over a search.

Table 27: **Runtime and compute.** Upper: measured fit cost by family over the runs behind the reported results — 13,277 fits and 189 wall-clock hours in total, a third of them tuned. A fit is one fold under S2 and one cell under S1 and S3, and its time covers the hyperparameter search, every seed refit and the evaluation; cells later withheld by the reporting floor are counted, because their fits were paid for. *Tuned* is the share of fits that carried a hyperparameter search, which is 0% for the pretrained tabular family by design. Lower: fit cost of the four pretrained models. TabICL and TabDPT have 204 fewer fits than TabPFN and TabFM-Torch because they were not run on the momentary extended targets (34 S1 cells and 34 five-fold S2 cells); on the core targets the four have the same fits at every timescale and setting.

By family

Family	Models	Fits	Tuned	Median s	p90 s	Max s	ms/row	Total h
Baseline	5	3,165	42%	2.5	46	1657	0.30	20.2
Deep Tabular	5	3,165	42%	12.7	223	3484	1.09	70.5
TFM	4	2,124	0%	15.0	108	707	1.62	26.5
DG	7	2,597	37%	7.4	134	9056	0.97	51.5
UDA	6	2,226	37%	6.1	92	1068	0.65	20.5
All	27	13,277	34%	7.1	117	9056	0.91	189.1

Pretrained tabular models

Measure	TabPFN	TabFM-Torch	TabICL	TabDPT
fits measured	633	633	429	429
median fit (s)	15.5	44.1	6.5	4.1
slowest fit (s)	649	707	527	590
largest training slice (rows)	171,018	171,018	171,018	171,018

B.7 Deviations from the reference protocol

We document four departures from, or implementation choices relative to, the reference protocol (Zhang et al., 2026b) to clarify how the reported experiments were conducted.

First, the 30-day cap on each participant's S1 history is applied only to the momentary track. The daily and weekly tracks use the available history because the same calendar-time cap would retain substantially different numbers of observations across tracks.

Second, S2 hyperparameter selection is performed independently within each outer fold. The outer training participants are further divided into participant-disjoint training and validation subsets. Each tunable model receives 30 Optuna trials, with hyperparameters selected by validation AUROC. The selected configuration is then used for the three seed runs within that fold. All fitted preprocessing statistics are estimated from the corresponding training subset, and outer-fold test participants are excluded from preprocessing, hyperparameter selection, model fitting, and checkpoint selection. Hyperparameters are not reused across outer folds.

Third, pretrained tabular models are evaluated without task-specific hyperparameter search.

Fourth, structural zero-filling precedes standardisation and is restricted to features for which the absence of an event represents a measured zero. Filling these values with zero after standardisation would instead encode the fitted mean: the participant-specific training mean under S1 or the pooled training mean under S2 and S3. Structural zero-filling uses predefined feature semantics rather than fitted statistics; subsequent preprocessing statistics are estimated from the training split only.

C Extended Results

C.1 S1 — within-user temporal prediction

The complete per-cell matrices hold 9,396 rows of (setting, track, feature space, dataset or direction, target, model, AUROC). Printed at body size they would run past a hundred pages, so the tables below report the aggregate views and the full matrices are released with the code.

Table 29: **Setting 1 results, by track, feature space and target.** Within-user temporal prediction, with one row per block and target. *Prob.* counts the dataset–target problems represented by each block, and *Cells* counts the corresponding model runs; dividing cells by problems therefore gives the number of models reporting in that block. All fourteen supervised models report in the main and screen-on blocks and twelve report in the extended blocks, where TabICL and TabDPT were not run. *Mean* and *Range* summarize pooled AUROC across those cells, and *Best* identifies the model with the highest mean in the block. Results are not averaged across targets because the targets represent distinct prediction problems. The three screen-on blocks report the sensed target of Appendix A.4, in the native space only; the top of the weekly range is StudentLife, whose S1 test set is 41 rows from eight participants, where two models reach 1.000. Per-cell participant counts, training and test rows, and class prevalence are reported in Table 12.

Target	Prob.	Cells	Mean	Range	Best
Momentary, shared					
stress	4	56	0.651	0.458–0.782	TabFM-T.
valence	3	42	0.614	0.452–0.730	TabFM-T.
arousal	2	28	0.616	0.523–0.745	TabFM-T.
Momentary, native					

Target	Prob.	Cells	Mean	Range	Best
stress	4	56	0.646	0.442–0.786	TabFM-T.
valence	3	42	0.618	0.468–0.729	TabFM-T.
arousal	2	28	0.620	0.512–0.747	TabFM-T.
Daily, shared					
stress	4	56	0.609	0.361–0.837	TabFM-T.
valence	3	42	0.584	0.389–0.861	TabFM-T.
arousal	2	28	0.540	0.455–0.726	TabFM-T.
Daily, native					
stress	5	70	0.599	0.413–0.847	TabFM-T.
valence	3	42	0.577	0.375–0.895	RF
arousal	2	28	0.542	0.394–0.728	TabFM-T.
Weekly, shared					
stress	3	42	0.563	0.482–0.656	TabICL
valence	3	42	0.564	0.463–0.638	TabTransf.
arousal	2	28	0.544	0.476–0.600	TabFM-T.
Weekly, native					
stress	3	42	0.580	0.430–0.669	TabFM-T.
valence	3	42	0.553	0.438–0.670	DCN
arousal	2	28	0.555	0.416–0.659	TabFM-T.
Weekly, longer-horizon instruments, native — GLOBEM and CES					
stress	1	14	0.583	0.505–0.708	TabFM-T.
dep	1	14	0.513	0.502–0.532	FT-Transf.
phq4	1	13	0.497	0.480–0.517	TabDPT
Momentary extended, shared — EmoPhone					
angry	1	12	0.692	0.503–0.824	TabFM-T.
anxious	1	12	0.709	0.559–0.827	TabFM-T.
cheerful	1	12	0.657	0.567–0.743	TabFM-T.
content	1	12	0.649	0.574–0.730	TabFM-T.
depressed	1	12	0.693	0.539–0.808	TabFM-T.
happy	1	12	0.670	0.549–0.767	TabFM-T.
relaxed	1	12	0.641	0.504–0.758	TabFM-T.
sad	1	12	0.697	0.545–0.824	TabFM-T.
attention	1	12	0.630	0.526–0.735	TabFM-T.
disturbance	1	12	0.670	0.544–0.784	TabFM-T.
mental	1	12	0.695	0.584–0.792	TabFM-T.
change	1	12	0.646	0.482–0.782	TabFM-T.
arousal_change	1	12	0.690	0.551–0.789	TabPFN
valence_change	1	12	0.644	0.477–0.766	TabFM-T.
Momentary extended, shared — K-EmoPhone					
attention	1	12	0.572	0.524–0.654	TabFM-T.
disturbance	1	12	0.697	0.539–0.828	TabFM-T.
change	1	12	0.625	0.468–0.769	TabPFN
Momentary extended, native — EmoPhone					
angry	1	12	0.682	0.476–0.824	TabFM-T.
anxious	1	12	0.711	0.557–0.827	TabFM-T.
cheerful	1	12	0.658	0.577–0.743	TabFM-T.
content	1	12	0.644	0.565–0.730	TabFM-T.
depressed	1	12	0.694	0.518–0.808	TabFM-T.
happy	1	12	0.671	0.566–0.768	TabFM-T.
relaxed	1	12	0.637	0.493–0.759	TabFM-T.
sad	1	12	0.693	0.507–0.824	TabFM-T.
disturbance	1	12	0.674	0.552–0.785	TabFM-T.
mental	1	12	0.699	0.576–0.792	TabFM-T.
change	1	12	0.656	0.504–0.781	TabFM-T.
arousal_change	1	12	0.692	0.540–0.790	TabPFN

Target	Prob.	Cells	Mean	Range	Best
valence_change	1	12	0.650	0.518–0.766	TabFM-T.
attention	1	12	0.644	0.539–0.739	RF
Momentary extended, native — K-EmoPhone					
attention	1	12	0.565	0.507–0.651	TabFM-T.
disturbance	1	12	0.690	0.513–0.828	TabFM-T.
change	1	12	0.638	0.508–0.767	TabPFN
Screen-on, momentary, native					
screen-on	4	56	0.721	0.615–0.810	TabFM-T.
Screen-on, daily, native					
screen-on	5	70	0.576	0.440–0.748	TabFM-T.
Screen-on, weekly, native					
screen-on	5	70	0.672	0.479–1.000	XGBoost

Table 30: **Setting 1 model ranks and levels.** Upper: the common roster, ranked over all 54 affective S1 problems of Table 29 (the six main blocks and CES weekly stress), in every one of which all fourteen supervised models report; *rank* orders models by mean AUROC and *mean rank* averages their per-block ranks, with the three timescale columns giving that average within each track. Lower: the mean AUROC each model reaches in each block. TabICL and TabDPT have no extended-track cells, marked –.

Model	Family	n	mean	rank	mean rank	mom.	daily	weekly
Common roster, all settings								
TabFM-Torch	TFM	54	0.673	1	2.58	1.0	3.2	3.6
TabICL	TFM	54	0.653	2	4.04	3.2	3.6	5.4
TabPFN	TFM	54	0.649	3	4.44	3.4	4.6	5.4
RF	Baseline	54	0.643	4	4.61	4.1	4.6	5.2
LightGBM	Baseline	54	0.630	5	5.58	4.5	5.4	6.9
XGBoost	Baseline	54	0.628	6	5.65	5.1	6.2	5.6
TabDPT	TFM	54	0.584	7	7.89	8.0	6.8	8.9
TabTransformer	Deep tab.	54	0.566	8	8.72	8.8	9.2	8.1
FT-Transformer	Deep tab.	54	0.554	10	9.30	8.4	9.6	9.8
ResNet	Baseline	54	0.557	9	9.46	11.2	10.4	6.6
MLP	Baseline	54	0.550	11	10.13	11.4	10.5	8.4
DCN	Deep tab.	54	0.549	12	10.28	11.7	11.1	7.9
SAINT	Deep tab.	54	0.534	13	10.72	11.1	10.5	10.6
TabNet	Deep tab.	54	0.526	14	11.59	13.1	9.3	12.6

Model	Mom. sh.	Mom. nat.	Day sh.	Day nat.	Wk sh.	Wk nat.	Wk track nat.	Ext. sh.
Levels within each block								
RF	0.701	0.707	0.645	0.632	0.578	0.584	0.555	0.757
XGBoost	0.695	0.702	0.601	0.606	0.560	0.593	0.563	0.755
LightGBM	0.696	0.703	0.637	0.603	0.554	0.572	0.560	0.759
MLP	0.572	0.559	0.524	0.540	0.562	0.548	0.512	0.575
ResNet	0.568	0.561	0.561	0.524	0.569	0.567	0.503	0.569
TabNet	0.530	0.538	0.537	0.526	0.516	0.509	0.502	0.542
DCN	0.561	0.558	0.537	0.519	0.567	0.560	0.501	0.564
SAINT	0.552	0.546	0.500	0.527	0.539	0.542	0.510	0.574
TabTransformer	0.591	0.595	0.544	0.546	0.564	0.561	0.517	0.660
FT-Transformer	0.587	0.598	0.527	0.546	0.526	0.538	0.519	0.668
TabPFN	0.708	0.709	0.648	0.650	0.578	0.583	0.552	0.763
TabFM-Torch	0.729	0.732	0.687	0.681	0.578	0.606	0.581	0.775
TabICL	0.710	0.715	0.658	0.658	0.576	0.583	0.542	–

Model	Mom. sh.	Mom. nat.	Day sh.	Day nat.	Wk sh.	Wk nat.	Wk track nat.	Ext. sh.
TabDPT	0.630	0.609	0.587	0.576	0.559	0.543	0.523	–

C.2 S2 — UNSEEN USERS WITHIN A STUDY

Whole participants are held out across five folds, grouped by participant and stratified by the label. The setting runs three rosters, reported as three blocks rather than as one ranking: fourteen supervised models, seven domain-generalization runs comprising six methods and their ERM control, and six adaptation methods against a matched source-only control. A shift-aware method is read as a difference from its own control, and each control shares the method's encoder, optimiser, schedule, batch size, early stopping and seeds, so the margin between them reflects the objective alone.

Table 31: **Setting 2 results, by model family.** Unseen users within a study, summarised as in Table 29. The three blocks are the setting's three rosters: the supervised models, the domain-generalization methods with their ERM control, and the adaptation methods with their source-only MLP control. Each control is listed inside its own block, so a control appearing as *Best* means that no method in that roster beat it. A shift-aware row is read as a difference from its own control and never against the supervised roster; the paired differences are in Tables 34 and 35. Each roster closes with three screen-on blocks, the sensed target of Appendix A.4, which runs in the native space only. Per-cell counts and prevalence are in Table 12.

Target	Prob.	Cells	Mean	Range	Best
Baselines, deep tabular and pretrained tabular					
Momentary, shared					
stress	4	56	0.538	0.492–0.612	TabPFN
valence	3	42	0.537	0.495–0.585	TabFM-T.
arousal	2	28	0.561	0.535–0.580	MLP
Momentary, native					
stress	4	56	0.535	0.475–0.584	TabFM-T.
valence	3	42	0.541	0.489–0.581	TabFM-T.
arousal	2	28	0.555	0.523–0.576	MLP
Daily, shared					
stress	4	56	0.542	0.457–0.597	TabTransf.
valence	3	42	0.515	0.441–0.592	MLP
arousal	2	28	0.538	0.503–0.573	DCN
Daily, native					
stress	5	70	0.534	0.483–0.612	TabTransf.
valence	3	42	0.511	0.420–0.587	TabTransf.
arousal	2	28	0.529	0.478–0.574	TabTransf.
Weekly, shared					
stress	3	42	0.494	0.419–0.529	FT-Transf.
valence	3	42	0.512	0.437–0.581	DCN
arousal	2	28	0.515	0.443–0.568	TabNet
Weekly, native					
stress	3	42	0.489	0.391–0.532	ResNet
valence	3	42	0.508	0.453–0.561	ResNet
arousal	2	28	0.513	0.441–0.575	XGBoost
Weekly, longer-horizon instruments, native — GLOBEM and CES					
stress	1	14	0.519	0.491–0.536	TabDPT
dep	1	14	0.527	0.500–0.548	TabICL
phq4	1	13	0.535	0.504–0.574	TabPFN
Momentary extended, shared — EmoPhone					
angry	1	12	0.516	0.498–0.578	TabPFN
anxious	1	12	0.540	0.521–0.570	RF

Target	Prob.	Cells	Mean	Range	Best
cheerful	1	12	0.565	0.504–0.589	MLP
content	1	12	0.549	0.521–0.566	TabFM-T.
depressed	1	12	0.535	0.480–0.586	TabPFN
happy	1	12	0.541	0.487–0.570	MLP
relaxed	1	12	0.503	0.440–0.534	MLP
sad	1	12	0.505	0.444–0.560	TabFM-T.
attention	1	12	0.541	0.509–0.561	XGBoost
disturbance	1	12	0.528	0.496–0.542	MLP
mental	1	12	0.558	0.507–0.591	TabTransf.
change	1	12	0.515	0.501–0.534	TabTransf.
arousal_change	1	12	0.534	0.509–0.566	RF
valence_change	1	12	0.539	0.488–0.567	RF
Momentary extended, shared — K-EmoPhone					
attention	1	12	0.542	0.527–0.569	MLP
disturbance	1	12	0.530	0.497–0.570	TabTransf.
change	1	12	0.503	0.470–0.525	ResNet
Momentary extended, native — EmoPhone					
angry	1	12	0.513	0.492–0.569	TabPFN
anxious	1	12	0.540	0.519–0.553	TabFM-T.
cheerful	1	12	0.566	0.529–0.589	TabTransf.
content	1	12	0.549	0.512–0.568	TabFM-T.
depressed	1	12	0.535	0.500–0.580	TabFM-T.
happy	1	12	0.545	0.514–0.569	MLP
relaxed	1	12	0.508	0.467–0.541	MLP
sad	1	12	0.507	0.466–0.565	TabFM-T.
attention	1	12	0.542	0.515–0.564	XGBoost
disturbance	1	12	0.523	0.492–0.540	MLP
mental	1	12	0.557	0.502–0.599	TabTransf.
change	1	12	0.512	0.482–0.542	SAINT
arousal_change	1	12	0.533	0.499–0.565	RF
valence_change	1	12	0.541	0.496–0.563	XGBoost
Momentary extended, native — K-EmoPhone					
attention	1	12	0.544	0.518–0.565	MLP
disturbance	1	12	0.528	0.494–0.563	RF
change	1	12	0.508	0.475–0.541	XGBoost
Screen-on, momentary, native					
screen-on	4	56	0.715	0.614–0.799	TabFM-T.
Screen-on, daily, native					
screen-on	5	70	0.563	0.468–0.697	RF
Screen-on, weekly, native					
screen-on	5	70	0.627	0.434–0.795	TabPFN

Target	Prob.	Cells	Mean	Range	Best
Domain generalization, including the ERM control					
Momentary, shared					
stress	4	28	0.534	0.464–0.561	Fish
valence	3	21	0.536	0.508–0.558	CSD
arousal	2	14	0.568	0.560–0.580	Fish
Momentary, native					
stress	4	28	0.532	0.495–0.565	VREx
valence	3	21	0.537	0.511–0.561	MixStyle
arousal	2	14	0.566	0.557–0.577	MixStyle
Daily, shared					
stress	4	28	0.535	0.478–0.594	ERM
valence	3	21	0.521	0.468–0.553	CSD

Target	Prob.	Cells	Mean	Range	Best
arousal	2	14	0.533	0.503–0.558	GroupDRO
Daily, native					
stress	5	35	0.530	0.477–0.595	GroupDRO
valence	3	21	0.527	0.491–0.557	ERM
arousal	2	14	0.529	0.498–0.550	MixStyle
Weekly, shared					
stress	3	21	0.509	0.492–0.538	MixStyle
valence	3	21	0.521	0.479–0.567	ERM
arousal	2	14	0.517	0.479–0.544	CSD
Weekly, native					
stress	3	21	0.516	0.508–0.537	Fish
valence	3	21	0.513	0.466–0.550	VREx
arousal	2	14	0.514	0.478–0.531	VREx
Weekly, longer-horizon instruments, native — GLOBEM and CES					
stress	1	7	0.527	0.526–0.529	ERM
dep	1	7	0.521	0.517–0.524	IRM
phq4	1	6	0.521	0.518–0.523	GroupDRO
Screen-on, momentary, native					
screen-on	4	28	0.705	0.535–0.785	Fish
Screen-on, daily, native					
screen-on	5	35	0.561	0.494–0.696	GroupDRO
Screen-on, weekly, native					
screen-on	5	35	0.616	0.438–0.781	MixStyle

Target	Prob.	Cells	Mean	Range	Best
Unsupervised domain adaptation, including the source-only MLP control					
Momentary, shared					
stress	4	28	0.533	0.501–0.561	JAN
valence	3	21	0.545	0.503–0.563	MCC
arousal	2	14	0.569	0.547–0.581	MLP
Momentary, native					
stress	4	28	0.532	0.499–0.562	MLP
valence	3	21	0.546	0.506–0.565	CORAL
arousal	2	14	0.567	0.548–0.577	DAN
Daily, shared					
stress	4	28	0.545	0.487–0.596	MCC
valence	3	21	0.538	0.475–0.579	CORAL
arousal	2	14	0.539	0.492–0.566	MLP
Daily, native					
stress	5	35	0.542	0.493–0.592	CDAN
valence	3	21	0.528	0.472–0.564	CDAN
arousal	2	14	0.544	0.515–0.563	MCC
Weekly, shared					
stress	3	21	0.504	0.474–0.519	JAN
valence	3	21	0.524	0.489–0.560	MCC
arousal	2	14	0.525	0.496–0.544	MLP
Weekly, native					
stress	3	21	0.508	0.472–0.527	MCC
valence	3	21	0.512	0.458–0.543	DAN
arousal	2	14	0.513	0.488–0.530	DAN
Weekly, longer-horizon instruments, native — GLOBEM and CES					
stress	1	7	0.524	0.500–0.531	DAN

Target	Prob.	Cells	Mean	Range	Best
dep	1	7	0.519	0.502–0.530	CDAN
phq4	1	6	0.518	0.512–0.521	DAN
Screen-on, momentary, native					
screen-on	4	28	0.712	0.624–0.788	MLP
Screen-on, daily, native					
screen-on	5	35	0.564	0.495–0.695	CDAN
Screen-on, weekly, native					
screen-on	5	35	0.637	0.483–0.782	DAN

C.3 S3 — CROSS-DATASET GENERALIZATION

One study is held out entire. Directions are enumerated as every ordered source–target pair together with the remaining datasets pooled against each held-out one. This produces 27 (direction, target) triples per timescale, which reduce to 23 distinct directions once the dataset pairs shared between targets are counted once (Appendix B.2). Every S3 cell is in the shared feature space, since a transfer direction has no meaning over columns one study does not collect. An S3 number therefore reflects both distribution shift and the constraint of cross-study feature alignment.

Table 32: **Setting 3 results, by transfer direction kind.** Cross-dataset generalization, summarised as in Table 29. *Prob.* counts (direction, target) triples rather than datasets. Over the affective targets they sum to 27 per timescale, which is the 23 distinct directions of Table 18 with the pairs serving more than one target counted once each; screen-on, the sensed target of Appendix A.4, adds nine at the momentary and daily timescales and two at the weekly one, all on directions of the valence roster. *Cells* counts the model runs behind a row, 27 per triple. *Rows* is the size of the held-out study's test set, which is a property of the target and so is shared by the single-source and pooled rows of a target. Arousal, and screen-on at the weekly timescale, have no pooled row because their roster is two datasets, where pooling degenerates to the pairwise direction and the duplicate is dropped. Every S3 cell is in the shared feature space. The affective means sit within about 0.04 AUROC of chance, so on those rows *Best* identifies the highest cell mean and should not be read as a stable ranking; screen-on transfers above chance at the momentary and weekly timescales, but a persistence rule reaches most of that (Table 33).

Target	Prob.	Cells	Rows	Mean	Range	Best
Momentary						
Single source						
stress	12	324	6,520	0.515	0.403–0.573	MLP
valence	6	162	51,941	0.504	0.463–0.539	TabFM-T.
arousal	2	54	28,661	0.541	0.509–0.565	DAN
Multiple sources						
stress	4	108	6,520	0.527	0.471–0.572	IRM
valence	3	81	51,941	0.510	0.480–0.536	VREx
Screen-on, single source						
screen-on	6	162	51,116	0.657	0.551–0.752	TabPFN
Screen-on, multiple sources						
screen-on	3	81	51,116	0.669	0.633–0.743	TabFM-T.
Daily						
Single source						
stress	12	324	1,014	0.522	0.385–0.621	RF
valence	6	162	5,769	0.496	0.414–0.567	TabNet
arousal	2	54	3,804	0.514	0.474–0.552	TabTransf.
Multiple sources						
stress	4	108	1,014	0.530	0.427–0.608	MLP

Target	Prob.	Cells	Rows	Mean	Range	Best
valence	3	81	5,769	0.500	0.441–0.571	DAN
Screen-on, single source						
screen-on	6	162	5,490	0.506	0.464–0.561	JAN
Screen-on, multiple sources						
screen-on	3	81	5,490	0.506	0.436–0.558	TabFM-T.
Weekly						
Single source						
stress	12	324	6,256	0.493	0.353–0.652	TabDPT
valence	6	162	40,575	0.501	0.444–0.564	MCC
arousal	2	54	20,996	0.508	0.458–0.564	ResNet
Multiple sources						
stress	4	108	6,256	0.493	0.345–0.577	RF
valence	3	81	40,575	0.501	0.441–0.554	CDAN
Screen-on, single source						
screen-on	2	54	67,662	0.604	0.570–0.640	TabFM-T.

Screen-on under S3. Three properties of the screen-on target bound what its S3 numbers can show. First, the label is person-centred: it is split at each participant's own median (Appendix A.4), computed on the training slice under S1 and over each participant's full series under S2 and S3, so every participant is close to half positive at every setting and the label carries no between-participant difference in prevalence. The affective labels are split at a fixed scale midpoint and their prevalence varies across participants and studies (Table 12); the two targets therefore pose different ranking problems, and levels should be compared within a target across settings rather than between the two targets at one setting. Second, the S3 roster is narrower than the S1 roster (Table 11): StudentLife is excluded at every timescale because its release keeps only episodes of at least an hour, so its reconstructed screen-on time is an upper bound of 12.6 hours a day against 4.0–6.6 hours elsewhere, a different quantity rather than a shifted one; CES is excluded at the daily timescale because the space it shares with the other studies is nine columns, seven of them screen features, and at the weekly one CES and GLOBEM share no columns with the rolled studies. The roster that remains is K-EmoPhone, EmoPhone and DiversityOne, and K-EmoPhone has no weekly screen-on label, so S3 has nine directions at the momentary and daily timescales and two at the weekly one (Table 18). Third, the features include the screen-on time of the window that precedes the label window.

Table 33 therefore sets the models against a rule that carries that preceding value forward, scored on the same test rows. At the momentary timescale the rule reaches 0.632 raw and 0.638 person-centred, against 0.667 for the fourteen supervised models: the models beat it clearly only when DiversityOne is held out (0.718 against 0.642), while on EmoPhone the rule is marginally ahead (0.649 against 0.640) and on K-EmoPhone behind by 0.022. At the daily timescale the rule is ahead (0.521 against 0.507). At the weekly timescale it falls below chance, 0.457 raw and 0.379 centred, because a participant's screen time in one week predicts a lower rather than a higher next week relative to their own median; the models' 0.604 is roughly the reversed person-centred rule, 0.621, so the weekly signal is this reversion toward the participant's median rather than a transferable behavioural pattern, and it rests on two directions. The screen-on contrast with affect is therefore real but narrow: the target transfers above chance mostly because a person's recent screen use carries over, which a model trained on another study can pick up from the same feature.

Table 33: **Screen-on under S3, against a persistence rule.** One row per held-out study and timescale, scored on exactly the test rows the models were scored on. *Dir.* counts the transfer directions into that study. *Raw* ranks test rows by the screen-on time of their own feature window, the quantity the label measures one window later; *Centred* subtracts the participant's median of it first, which matches how the label is split. Neither rule is fitted to anything. *14 models* is the mean over the supervised roster that the main text reports, *27 models* the mean over every model and *Best* the best model's mean over the directions into that study. A persistence AUROC below 0.5 is reported as measured: reversing its sign would be a fitted choice.

			Persistence		Models		
Held-out study	Dir.	Test rows	Raw	Centred	14 models	27 models	Best
Momentary							
DiversityOne	3	160,546	0.640	0.642	0.718	0.707	0.737
EmoPhone	3	51,116	0.639	0.649	0.640	0.636	0.653
K-EmoPhone	3	5,305	0.617	0.622	0.644	0.640	0.654
Daily							
DiversityOne	3	5,490	0.530	0.545	0.517	0.518	0.530
EmoPhone	3	7,593	0.541	0.556	0.513	0.510	0.536
K-EmoPhone	3	463	0.491	0.460	0.491	0.492	0.513
Weekly							
DiversityOne	1	99,464	0.462	0.360	0.611	0.612	0.637
EmoPhone	1	35,859	0.452	0.399	0.596	0.595	0.640

C.4 Generalization gaps across configurations

The distribution of generalization gaps matters more than their mean because the aggregate ordering of the settings is not reproduced by every individual configuration. Under pooled AUROC, the cross-user step from S1 to S2 is negative in 77% of matched configurations, while the cross-study step from S2 to S3 is negative in 73%. The pooled S1-to-S2 differences have an interquartile range from -0.111 to -0.003 AUROC.

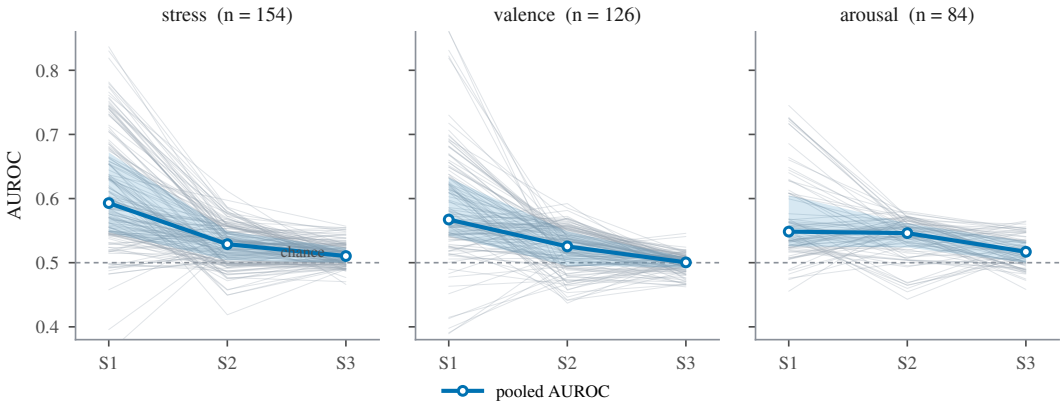


Figure 10: **Per-configuration generalization trajectories.** Each line represents one matched configuration across the three evaluation settings. The trajectories show variation among configurations that is not visible from aggregate means alone.

The spread between configurations also narrows as the setting becomes more challenging. The interquartile range decreases from 0.0981 AUROC under S1 to 0.0490 under S2 and 0.0221 under S3, while the proportion of configurations scoring above chance falls from 92.0% to 78.8% and then to 66.5%. Cross-study evaluation therefore does more than lower performance; it compresses the field toward chance, leaving little room for models to separate under S3 and producing the unstable S3 rankings reported in Appendix C.8.

C.5 Where the setting effects vary

Transfer asymmetry. Forward and reverse transfer differ substantially in size, but not reliably in direction. Averaged within each dataset pair, the mean absolute difference between A→B and B→A ranges from 0.021 to 0.045 AUROC (0.020 to 0.033 when averaged within each timescale instead, Table 36) across the dataset pairs, which is larger than the whole cross-dataset step. The signed means, meanwhile, sit within ± 0.011 of zero, and the share of positive differences within a pair ranges from 33% to 68%. Transfer asymmetry is therefore a property of individual configurations

rather than a stable ordering of datasets, and no dataset here should be described as a better source than another.

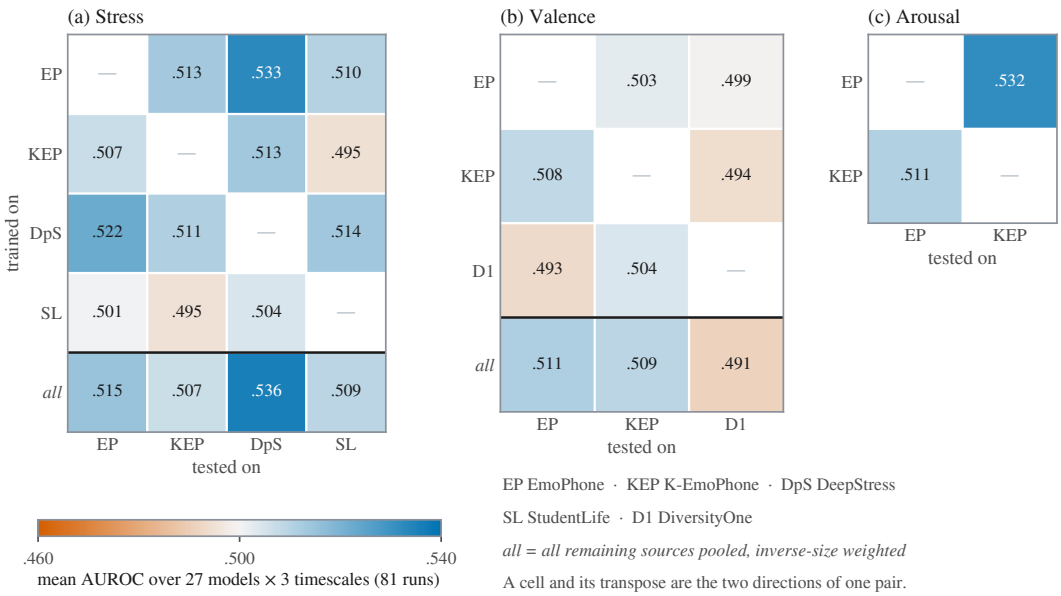


Figure 11: **Source-to-target transfer, every direction.** Rows are the training dataset and columns the held-out test dataset; each cell averages that direction's model×timescale runs. A cell read against its transpose is one pair's asymmetry.

One panel per target is unnecessary. The pattern holds on all three affective targets: a cell and its transpose differ in magnitude but not reliably in direction. The per-target matrices are released with the code.

Single against multiple sources. Pooling every available source does not match the best single source. One caveat belongs with the numbers before they are read: pooled directions are trained with inverse-size sample weighting while single-source directions are unweighted, so this comparison cannot separate the effect of adding data from the effect of balancing it. Against the best single source for each problem, training on all remaining datasets loses 0.0145 AUROC over 567 matched model comparisons, wins in 25% of them, and wins on one of seven problems. The deficit holds at all three timescales and widens as the window lengthens, from -0.0097 at momentary to -0.0184 at weekly.

C.6 Model families, and the regime the pretrained models face

The baseline family is not a set of weak controls. Tree ensembles are strong on structured sensing features, and which baseline serves as the practical reference point depends on the setting. We therefore compare the pretrained models against two baselines: one fixed in advance, and the strongest baseline in each cell. The two comparators are biased in opposite directions, and neither settles the question alone.

The pretrained tabular family shows the benchmark's sharpest interaction. It holds the highest family mean under S1 at every timescale and the lowest or near-lowest under S3 at momentary and daily, and it takes the most outright cell wins per member at S1 throughout. A family that leads on average and wins most individual cells at one setting, while doing neither at another, is not uniformly strong or weak. It is strong where the evaluation leaves a person's own history available, and not where the study changes.

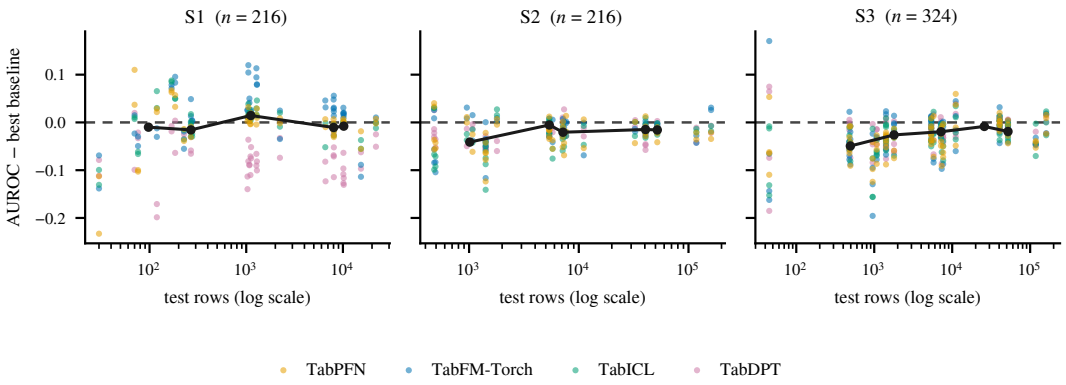


Figure 12: **Pretrained tabular models against training-set size.** Relative performance by the training rows a cell supplies. The context cap binds on the largest sources, so a deficit there cannot separate a weaker model from a starved one.

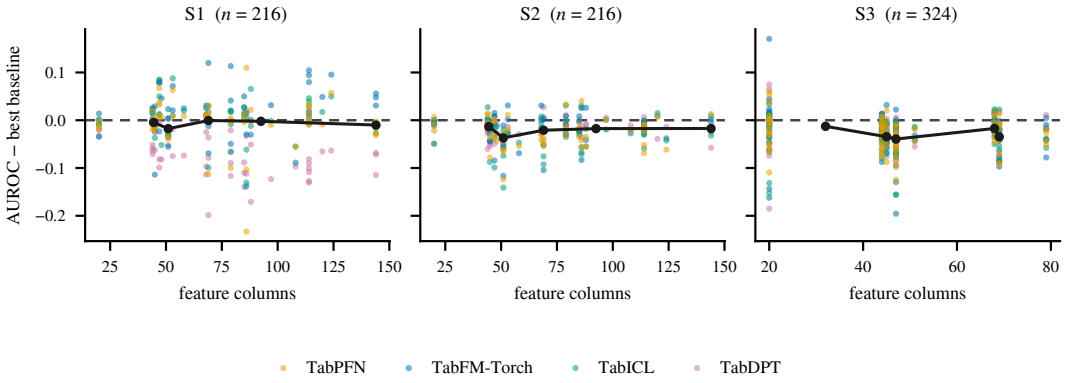


Figure 13: **Pretrained tabular models against feature width.** Relative performance by the number of columns a cell supplies, which the Native and Shared conditions vary by construction.

The context cap is what these figures require to be read correctly. TabFM-Torch caps its in-context set at 10,000 rows while TabICL uses the full training slice, so on the largest cross-dataset sources one model observes a small fraction of the context available to the other. A deficit on a large-source cell therefore cannot by itself distinguish a weaker model from a starved one. The sample-size figure describes the benchmark's data regime and is not an ablation of context length.

C.7 Shift-aware objectives against their controls

Both shift-aware families are evaluated as paired differences from a matched control on identical cells. This is the comparison the benchmark is built to answer: a method and its control share the baseline multilayer perceptron's encoder and every element of the training loop, and differ only in the loss term, so a margin reflects the objective rather than the architecture.

Table 34: **Domain generalization against its ERM control.** Each method is paired with the control on identical cells and differs from it only in the loss term, so a margin reflects the objective rather than the architecture. Equivalence is two one-sided tests at ±0.01 AUROC.

Setting	Method	Timescale	n	mean Δ	med. Δ	Δ>0	$\lvert\Delta\rvert \leq 0.01$
By track							
S2	CSD	Day	19	-0.0037	-0.0049	37%	53%
S2	CSD	Mom.	18	+0.0005	-0.0005	50%	89%
S2	CSD	Wk	17	-0.0014	-0.0004	47%	82%

Setting	Method	Timescale	n	mean Δ	med. Δ	Δ>0	$\lvert\Delta\rvert \leq 0.01$
S2	Fish	Day	19	-0.0018	-0.0019	47%	84%
S2	Fish	Mom.	18	+0.0011	+0.0002	50%	78%
S2	Fish	Wk	17	+0.0011	-0.0002	47%	71%
S2	GroupDRO	Day	19	+0.0012	-0.0002	42%	58%
S2	GroupDRO	Mom.	18	-0.0021	-0.0020	28%	83%
S2	GroupDRO	Wk	17	-0.0029	+0.0011	53%	71%
S2	IRM	Day	19	-0.0205	-0.0193	21%	37%
S2	IRM	Mom.	18	-0.0040	-0.0020	28%	72%
S2	IRM	Wk	17	-0.0038	-0.0020	41%	65%
S2	MixStyle	Day	19	-0.0003	-0.0014	37%	68%
S2	MixStyle	Mom.	18	+0.0010	+0.0002	56%	89%
S2	MixStyle	Wk	17	-0.0004	+0.0016	53%	76%
S2	VREx	Day	19	-0.0105	-0.0091	37%	42%
S2	VREx	Mom.	18	-0.0055	-0.0017	28%	72%
S2	VREx	Wk	17	-0.0037	-0.0035	41%	47%
S3	CSD	Day	27	+0.0085	+0.0086	78%	56%
S3	CSD	Mom.	27	-0.0024	-0.0016	41%	81%
S3	CSD	Wk	27	+0.0005	-0.0028	37%	59%
S3	Fish	Day	27	+0.0083	+0.0067	78%	70%
S3	Fish	Mom.	27	-0.0023	+0.0001	52%	85%
S3	Fish	Wk	27	+0.0003	-0.0013	37%	59%
S3	GroupDRO	Day	27	+0.0039	+0.0004	52%	52%
S3	GroupDRO	Mom.	27	-0.0019	-0.0001	48%	78%
S3	GroupDRO	Wk	27	+0.0037	+0.0002	59%	56%
S3	IRM	Day	27	+0.0054	+0.0053	56%	52%
S3	IRM	Mom.	27	-0.0047	-0.0029	41%	67%
S3	IRM	Wk	27	+0.0074	+0.0041	67%	59%
S3	MixStyle	Day	27	+0.0032	+0.0005	59%	63%
S3	MixStyle	Mom.	27	+0.0019	+0.0007	63%	81%
S3	MixStyle	Wk	27	-0.0007	-0.0027	41%	48%
S3	VREx	Day	27	-0.0034	-0.0003	48%	48%
S3	VREx	Mom.	27	-0.0046	-0.0034	41%	67%
S3	VREx	Wk	27	+0.0043	+0.0003	52%	56%

Setting	Method	n	mean Δ	med. Δ	95% CI	Holm p	d_z	win–loss
Paired tests (1 of 2)								
S2	MixStyle	54	+0.0001	-0.0003	[-0.0027, +0.0031]	1.00	+0.01	26–28
S2	Fish	54	+0.0001	-0.0001	[-0.0030, +0.0032]	1.00	+0.01	26–28
S2	GroupDRO	54	-0.0012	-0.0011	[-0.0050, +0.0025]	1.00	-0.08	22–32
S2	CSD	54	-0.0015	-0.0008	[-0.0050, +0.0019]	1.00	-0.12	24–30
S2	VREx	54	-0.0067	-0.0021	[-0.0124, -0.0012]	0.350	-0.32	19–35
S2	IRM	54	-0.0097	-0.0037	[-0.0157, -0.0041]	0.021	-0.44	16–38
S3	IRM	81	+0.0027	+0.0011	[-0.0018, +0.0073]	1.00	+0.13	44–37
S3	CSD	81	+0.0022	+0.0016	[-0.0016, +0.0064]	1.00	+0.12	42–39
S3	Fish	81	+0.0021	+0.0010	[-0.0009, +0.0051]	0.745	+0.15	45–36
S3	GroupDRO	81	+0.0019	+0.0001	[-0.0017, +0.0058]	1.00	+0.11	43–38
S3	MixStyle	81	+0.0015	+0.0004	[-0.0016, +0.0046]	1.00	+0.10	44–37
S3	VREx	81	-0.0012	-0.0010	[-0.0053, +0.0030]	1.00	-0.06	38–43

Setting	Method	equiv.
Paired tests (2 of 2)		
S2	MixStyle	yes
S2	Fish	yes
S2	GroupDRO	yes
S2	CSD	yes
S2	VREx	no
S2	IRM	no
S3	IRM	yes
S3	CSD	yes
S3	Fish	yes
S3	GroupDRO	yes
S3	MixStyle	yes
S3	VREx	yes

Table 35: **Unsupervised adaptation against a matched source-only model.** The comparator is a source-only model of identical capacity, not the strongest supervised model. These methods additionally observe the target's unlabelled test-split features, so a transductive method is being compared with a non-transductive control.

Setting	Method	Timescale	n	mean Δ	med. Δ	$\Delta>0$	$\lvert\Delta\rvert \leq 0.01$
By track							
S2	CDAN	Day	19	+0.0032	+0.0032	53%	53%
S2	CDAN	Mom.	18	-0.0016	+0.0005	61%	89%
S2	CDAN	Wk	17	-0.0045	-0.0039	29%	71%
S2	CORAL	Day	19	+0.0011	-0.0008	37%	53%
S2	CORAL	Mom.	18	-0.0001	+0.0014	56%	78%
S2	CORAL	Wk	17	+0.0008	+0.0009	59%	53%
S2	DAN	Day	19	-0.0028	-0.0072	42%	42%
S2	DAN	Mom.	18	-0.0004	+0.0022	61%	78%
S2	DAN	Wk	17	+0.0041	-0.0008	41%	53%
S2	DANN	Day	19	-0.0005	-0.0024	42%	47%
S2	DANN	Mom.	18	-0.0033	-0.0008	39%	78%
S2	DANN	Wk	17	+0.0009	+0.0004	59%	71%
S2	JAN	Day	19	-0.0148	-0.0110	16%	42%
S2	JAN	Mom.	18	-0.0055	-0.0059	28%	61%
S2	JAN	Wk	17	-0.0105	-0.0097	41%	53%
S2	MCC	Day	19	+0.0019	+0.0016	63%	58%
S2	MCC	Mom.	18	+0.0012	+0.0004	56%	89%
S2	MCC	Wk	17	+0.0060	+0.0015	76%	65%
S3	CDAN	Day	27	-0.0046	-0.0011	37%	59%
S3	CDAN	Mom.	27	-0.0048	-0.0019	30%	78%
S3	CDAN	Wk	27	+0.0002	+0.0044	59%	59%
S3	CORAL	Day	27	-0.0021	-0.0006	48%	59%
S3	CORAL	Mom.	27	-0.0032	-0.0036	30%	74%
S3	CORAL	Wk	27	+0.0011	+0.0024	63%	52%
S3	DAN	Day	27	-0.0024	+0.0003	52%	52%
S3	DAN	Mom.	27	-0.0035	-0.0002	48%	63%
S3	DAN	Wk	27	-0.0005	+0.0003	52%	52%
S3	DANN	Day	27	-0.0061	-0.0012	41%	63%
S3	DANN	Mom.	27	-0.0020	-0.0025	33%	81%
S3	DANN	Wk	27	-0.0070	-0.0058	30%	44%
S3	JAN	Day	27	-0.0052	-0.0048	30%	44%
S3	JAN	Mom.	27	-0.0040	-0.0025	41%	63%
S3	JAN	Wk	27	-0.0030	-0.0005	48%	30%
S3	MCC	Day	27	-0.0040	-0.0056	26%	56%

Setting	Method	Timescale	n	mean Δ	med. Δ	Δ>0	$\lvert\Delta\rvert \leq 0.01$
S3	MCC	Mom.	27	-0.0034	-0.0027	44%	56%
S3	MCC	Wk	27	-0.0030	-0.0008	48%	52%

Setting	Method	n	mean Δ	med. Δ	95% CI	Holm p	d_z	win–loss
Paired tests (1 of 2)								
S2	MCC	54	+0.0030	+0.0013	[-0.0010, +0.0069]	0.256	+0.20	35–19
S2	CORAL	54	+0.0006	-0.0000	[-0.0036, +0.0047]	1.00	+0.04	27–27
S2	DAN	54	+0.0001	-0.0002	[-0.0039, +0.0043]	1.00	+0.01	26–28
S2	CDAN	54	-0.0008	-0.0007	[-0.0048, +0.0032]	1.00	-0.06	26–28
S2	DANN	54	-0.0010	-0.0008	[-0.0051, +0.0033]	1.00	-0.07	25–29
S2	JAN	54	-0.0104	-0.0072	[-0.0149, -0.0059]	<0.001	-0.61	15–39
S3	CORAL	81	-0.0014	-0.0007	[-0.0045, +0.0016]	0.786	-0.10	38–43
S3	DAN	81	-0.0021	+0.0002	[-0.0061, +0.0018]	0.786	-0.12	41–40
S3	CDAN	81	-0.0031	-0.0005	[-0.0076, +0.0010]	0.732	-0.15	34–47
S3	MCC	81	-0.0035	-0.0033	[-0.0090, +0.0014]	0.732	-0.15	32–49
S3	JAN	81	-0.0041	-0.0032	[-0.0094, +0.0012]	0.732	-0.16	32–49
S3	DANN	81	-0.0050	-0.0027	[-0.0086, -0.0017]	0.019	-0.32	28–53

Setting	Method	equiv.
Paired tests (2 of 2)		
S2	MCC	yes
S2	CORAL	yes
S2	DAN	yes
S2	CDAN	yes
S2	DANN	yes
S2	JAN	no
S3	CORAL	yes
S3	DAN	yes
S3	CDAN	yes
S3	MCC	yes
S3	JAN	yes
S3	DANN	yes

Null results here are assessed by equivalence testing rather than by failure to reject. We run two one-sided tests against a margin of ± 0.01 AUROC, equivalently requiring the 90% interval of the mean difference to fall inside that margin. The margin is a judgement about practical relevance rather than a derived quantity: below one AUROC point, no deployment decision in this literature changes. The adaptation methods carry one further asymmetry wherever they appear. They observe the target's test-split feature matrix without labels, which is the exact set of rows on which they will be scored, so their comparison is between a transductive method and a non-transductive control.

C.8 Model ranks and their stability across settings

Ranks are computed on the common roster, meaning the models that report in every block being compared, since a rank taken over different rosters is not comparable.

Rank agreement is strong under S1, moderate under S2, and barely present under S3. This pattern reflects differences among the evaluation settings rather than stable differences among models. Under S3, the best and worst models are separated by only approximately 0.020 AUROC, making the resulting ranking highly sensitive to seed and fold variation. We therefore interpret rank correlations descriptively rather than as confirmatory evidence, because they are computed from overlapping benchmark tasks and do not have a clearly defined statistical null.

C.9 Extended RQ3 analyses — prediction timescale

All three evaluation settings are applied unchanged at each of the three prediction timescales, which makes this a replication question. The analyses above are re-run separately inside each timescale and compared for agreement rather than for level.

One comparability statement governs the whole section: the timescales do not run the same experiment. The tracks cover different datasets. A row is one prompt, one person-day or one person-week. On the three shared affective targets, the weekly label is derived from the same repeated self-reports the momentary track uses rather than from a separately administered weekly instrument. A raw AUROC difference between timescales therefore mixes the prediction problem, the population and the label construction, and we read no such difference as evidence that one timescale is intrinsically easier. What is comparable is the pattern inside a timescale, which is why every setting gap is reported three times rather than pooled.

Table 36: **The three prediction timescales, side by side.** Which combinations each timescale evaluates is Table 11, and which models report in them Table 19, so only the levels are given here. The timescales do not run the same experiment: the tracks cover different datasets, a row is a prompt, a person-day or a person-week, and the weekly affective label aggregates the momentary track's own responses. Only the pattern inside a timescale is comparable, which is why every setting gap is reported three times rather than pooled.

Timescale	Setting	mean	med.	>0.5	best model
Coverage					
Day	S1	0.583	0.557	84%	TabFM-Torch 0.684
Day	S2	0.531	0.534	85%	TabTransformer 0.549
Day	S3	0.514	0.514	71%	MLP 0.523
Mom.	S1	0.631	0.631	96%	TabFM-Torch 0.731
Mom.	S2	0.542	0.547	96%	TabFM-Torch 0.555
Mom.	S3	0.516	0.517	75%	MixStyle 0.523
Wk	S1	0.563	0.555	91%	TabFM-Torch 0.599
Wk	S2	0.510	0.510	68%	MCC 0.521
Wk	S3	0.497	0.499	46%	DCN 0.506

Timescale	Comp.	n	mean Δ	95% CI	scope
Setting gaps					
Day	S2 - S1	266	-0.0539	[-0.0663, -0.0417]	daily, within timescale
Day	S3 - S1	126	-0.0751	[-0.0936, -0.0574]	daily, within timescale
Day	S3 - S2	243	-0.0205	[-0.0245, -0.0166]	daily, within timescale
Mom.	S2 - S1	252	-0.0888	[-0.0990, -0.0791]	momentary, within timescale
Mom.	S3 - S1	126	-0.1154	[-0.1314, -0.1000]	momentary, within timescale
Mom.	S3 - S2	243	-0.0241	[-0.0270, -0.0212]	momentary, within timescale
Wk	S2 - S1	238	-0.0580	[-0.0663, -0.0500]	weekly, within timescale
Wk	S3 - S1	112	-0.0597	[-0.0679, -0.0516]	weekly, within timescale
Wk	S3 - S2	216	-0.0088	[-0.0132, -0.0043]	weekly, within timescale

Timescale	Family	S1	S2	S3	S1 rank	S2 rank	S3 rank
Model families							
Day	Baseline	0.587	0.528	0.516	7.4	14.9	14.0
Day	Deep Tab.	0.531	0.530	0.516	9.9	14.5	13.7
Day	DG	–	0.529	0.517	–	15.2	13.4
Day	TFM	0.643	0.529	0.500	4.5	14.0	17.6
Day	UDA	–	0.539	0.519	–	11.5	12.5
Mom.	Baseline	0.646	0.540	0.510	7.3	15.9	15.6
Mom.	Deep Tab.	0.566	0.540	0.517	10.6	14.8	14.0
Mom.	DG	–	0.542	0.519	–	14.0	12.5
Mom.	TFM	0.693	0.547	0.508	3.9	13.9	16.6
Mom.	UDA	–	0.544	0.519	–	11.8	12.6
Wk	Baseline	0.571	0.505	0.499	6.6	15.6	14.5
Wk	Deep Tab.	0.541	0.512	0.497	9.8	14.0	14.5
Wk	DG	–	0.516	0.496	–	11.8	13.4
Wk	TFM	0.580	0.496	0.497	5.8	18.0	14.5
Wk	UDA	–	0.514	0.496	–	12.5	13.5

Timescale	Family	Setting	methods	mean Δ	Δ>0	$\lvert\Delta\rvert \leq 0.01$
Shift-aware families						
Day	DG	S2	6	-0.0059	17%	57%
Day	DG	S3	6	+0.0043	83%	57%
Day	UDA	S2	6	-0.0020	50%	49%
Day	UDA	S3	6	-0.0041	0%	56%
Mom.	DG	S2	6	-0.0015	50%	81%
Mom.	DG	S3	6	-0.0023	17%	77%
Mom.	UDA	S2	6	-0.0016	17%	79%
Mom.	UDA	S3	6	-0.0035	0%	69%
Wk	DG	S2	6	-0.0018	17%	69%
Wk	DG	S3	6	+0.0026	83%	56%
Wk	UDA	S2	6	-0.0006	67%	61%
Wk	UDA	S3	6	-0.0020	33%	48%

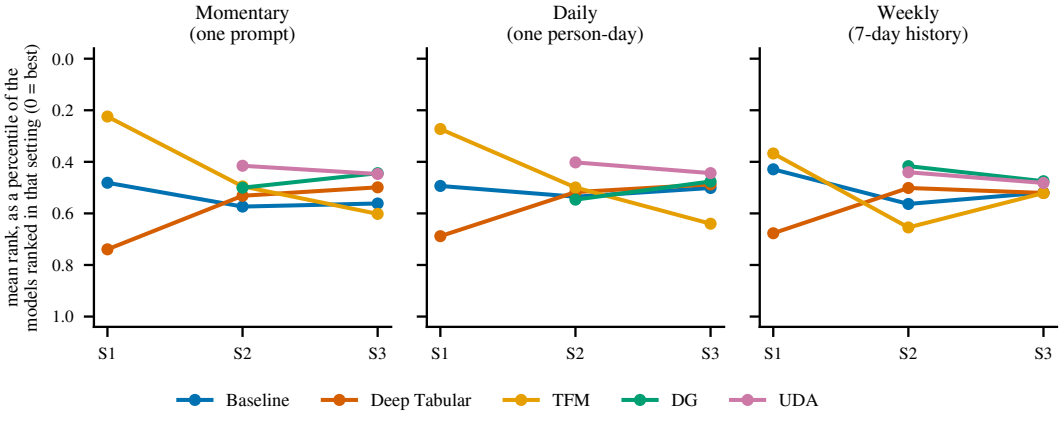


Figure 14: **Model-family ranks by prediction timescale.** Family-level ordering within each timescale and setting.

Timescale	Training	dir.	cells	mean	>0.5	best dir.	worst dir.	asym.
Cross-dataset transfer (Table 36)								

Timescale	Training	dir.	cells	mean	>0.5	best dir.	worst dir.	asym.
Day	multi-source	7	189	0.518	74%	0.552	0.486	–
Day	single-source	16	540	0.513	69%	0.559	0.486	0.033
Mom.	multi-source	7	189	0.520	86%	0.545	0.500	–
Mom.	single-source	16	540	0.514	71%	0.549	0.483	0.020
Wk	multi-source	7	189	0.496	43%	0.516	0.478	–
Wk	single-source	16	540	0.497	47%	0.523	0.462	0.032

Each timescale independently reproduces the pooled decline from S1 to S2 and from S2 to S3. The cross-user gap is negative at every timescale: -0.0888 momentary, -0.0539 daily, and -0.0580 weekly. S3 also remains below S2 at every timescale, with gaps of -0.0241 momentary, -0.0205 daily, and -0.0088 weekly.

The apparent reduction in the cross-dataset gap across timescales reflects a floor effect rather than a benefit of temporal aggregation. By the weekly timescale, the matched S2 and S3 means are at or near chance, leaving little remaining performance to lose. Aggregating observations over a seven-day window therefore does not improve cross-study transfer; instead, the available predictive signal is already weak before the cross-dataset shift is applied.