NOT ALL FORGETTING IS EQUAL: OBJECTIVE STRUCTURE IN TEMPORAL CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Catastrophic forgetting is often treated as a uniform cost of sequential updates, yet on structured temporal streams its severity can depend on the prediction objective. We study Domain-IL fine-tuning on three matched univariate corpora—glucose, electricity, and weather—under five objectives (next-value and short-/long-horizon regression; dense and rare within-horizon events) and five architectures. Event objectives forget far more than value objectives in every domain–architecture cell. Controls rule out rarity, horizon, and the monthly protocol; a reference swap shows the gap follows how the target is defined (entity-fixed thresholds vs. segment-relative value targets), not events as such. Joint-vs-sequential and recurrence diagnostics then split failures into preventable interference and a recurrence gap invisible to single-transition forgetting. Savings, recalibration, and mechanistic probes further show that the loss is often recoverable for recurrent learners, while transformer damage reaches the representation. We show that not all forgetting on streams is equal. Rather, it is objective-conditioned and diagnostically typed.

1 INTRODUCTION

Catastrophic forgetting remains a central obstacle to continual learning, as models that adapt to new data often lose performance on earlier tasks (Hoi et al., 2021; Du, 2025; Toneva et al., 2019; Kemker et al., 2018; Van de Ven et al., 2024). Most empirical work quantifies this with output-level metrics such as accuracy drop or backward transfer (BWT) (Lopez-Paz & Ranzato, 2017; Khodaee et al., 2024; Mirzadeh et al., 2022). These metrics are useful, but they treat every performance drop as the same kind of failure. They do not ask whether the loss depends on the prediction objective, how it fails, or whether it can be recovered (Díaz-Rodríguez et al., 2018; Khodaee et al., 2024).

That distinction matters for real sequential streams. In longitudinal monitoring (e.g., glucose, electricity load, weather) strong shared structure can persist across temporal regimes (Nestor et al., 2019; Aljundi et al., 2019; Kovatchev et al., 2009), so a model may show measurable degradation after a switch while still retaining a reusable solution (Martin et al., 2026; Nestor et al., 2019). Treating all drops as equally catastrophic can overstate failure when the objective is well aligned with that structure, and understate it when the objective is not.

We study Domain-IL sequential fine-tuning on three univariate corpora (glucose, electricity, weather), under five fixed objectives on the same streams: next-value and short-/long-horizon value regression, and dense and rare within-horizon event classification. Segments are monthly calendar blocks; learners are LSTM, GRU, RNN, CNN, and Transformer. We ask:

i. Does forgetting depend primarily on the objective, or on architecture and capacity (Q1)?
ii. Is event forgetting a single failure mode or do diagnostics split (Q2)?
iii. How recoverable is that loss, and where in the model does it sit (Q3)?

We find that event objectives forget far more than value objectives under a matched protocol (Q1). We see that this gap follows how the target is referenced rather than rarity or horizon (Q2). Lastly, we show that the damage is often recoverable for recurrent learners but representational for the transformer (Q3).

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.

2 Related work

2.1 Catastrophic forgetting and continual learning diagnostics

Classical continual learning evaluation emphasizes output-level degradation, typically through final-task performance (Lopez-Paz & Ranzato, 2017; Van de Ven et al., 2024; Aljundi et al., 2019). These metrics define the field's baseline language and remain essential for cross-paper comparability. However, they summarize *what* is lost in performance, and not necessarily *how* recoverable the learned solution remains. Our work follows the literature in reporting output-level forgetting, but argues that these metrics alone can conflate reversible perturbation with more damaging interference.

A broad line of continual learning work studies the stability-plasticity tradeoff using regularization, replay, architectural constraints, or parameter isolation (Jung et al., 2023; Lu et al., 2025; Kumar et al., 2023; Kirkpatrick et al., 2017). Our focus is narrower. Given the same sequential shift setting, how much variance in forgetting is attributable to objective versus architecture and domain? We position objective as a key modulator of forgetting and architecture as a modulator of vulnerability within an objective-dependent forgetting.

2.2 Temporal / online / concept-drift learning for time series

Continual learning has been fairly well explored, but most empirical evidence for catastrophic forgetting comes from vision based classification benchmarks (Lange et al., 2021; Mai et al., 2021; Wang et al., 2024) or from synthetic task sequences built with permutation and transformation on single datasets, most notably MNIST (Goodfellow et al., 2015; Kirkpatrick et al., 2017; van de Ven et al., 2022; LeCun et al., 2010). Forgetting in controlled settings where task similarity can be specified has also been studied with teacher-student networks and regression tasks (Lee et al., 2021; Lin et al., 2023). However, time series studies rarely compare how different objectives on the same stream forget under identical sequential updates.

2.3 Event detection and rare-event learning

Prior work (and NeurIPS workshops (e.g., 2021 Distribution Shifts (Sagawa et al.; Malinin et al., 2022))) often links forgetting and generalization failures to distributional shift (Guo et al., 2024; Nasery et al., 2021). Distance metrics, such as Wasserstein-based measures, provide a useful summary of shift severity, but their explanatory power can vary substantially by task and representation regime (Mallasto & Feragen, 2017). Our results align with this view; distance-forgetting association is weak on persistence-dominated next-value prediction and stronger on hypoglycemia classification, suggesting that shift metrics should be interpreted through task structure rather than treated as universally predictive.

2.4 Relearning and savings as a complementary signal

The notion of savings, specifically faster relearning after prior exposure, has deep roots in memory research and appears in machine learning as recovery speed or relearning efficiency (Ebbinghaus , 1885; Fellers & Storm, 2024; Tran & Woo, 2026; Kang et al., 2025). We adopt savings as a formal operational criterion, where positive savings indicates retained reusable structure, while negative savings indicates relearning impairment relative to learning from scratch. In our framing, this does not replace standard forgetting metrics; instead it complements them by disambiguating regimes with similar output degradation, but with different relearning dynamics.

2.5 Representation similarity / probing in continual learning.

Recent discussions in continual learning increasingly distinguish between output degradation from representation damage, but direct representation probing is not always available in applied settings (Davari et al., 2022; Lanza et al., 2026; Choudhuri et al., 2023; Aitken et al., 2022). Our approach is deliberately behavioral. Through our experiments on real world datasets, we correlate measured representational damage with negative savings and recovery dynamics.

3 PROBLEM DEFINITION

We model sequential fine-tuning on structured univariate streams as a Domain-Incremental Learning (van de Ven et al., 2022) problem: a *fixed* prediction objective is trained across a non-stationary sequence of temporal domains. Let $\{(x_t, \tau_t)\}_{t=1}^{T}$ be a univariate series ($x_t \in \mathbb{R}$). Operationally, we partition the stream into chronologically ordered *temporal segments* $\mathcal{S}_1, \ldots, \mathcal{S}_M$, each inducing a segment-specific distribution P_m. We interpret these segments as chronological proxies for latent *regimes* of a nonstationary process (e.g., time-series, HMM sense (Hamilton, 1989; Kim & Halbert; Rabiner, 1989)), exhibiting discrete operating conditions of the dynamics, not as continual learning "tasks" with changing label spaces or task identities. From sliding windows of length L we form

$$\mathcal{D}_m = \left\{ \left(x_{t_k - L + 1 : t_k},\, y_{m,k} \right) \right\}_{k=1}^{K},$$

where the target y is set by the objective under study (e.g., next-step regression or within-horizon event labels). A learner is updated sequentially $\mathcal{D}_1 \to \mathcal{D}_2 \to \cdots$; forgetting is the degradation of held-out performance on earlier *segments* after later updates. We repeat this protocol under several fixed objectives on the same stream, so Domain-IL (shifting P_m) and objective structure (different y) are separate axes of analysis, as shown in Figure 1.

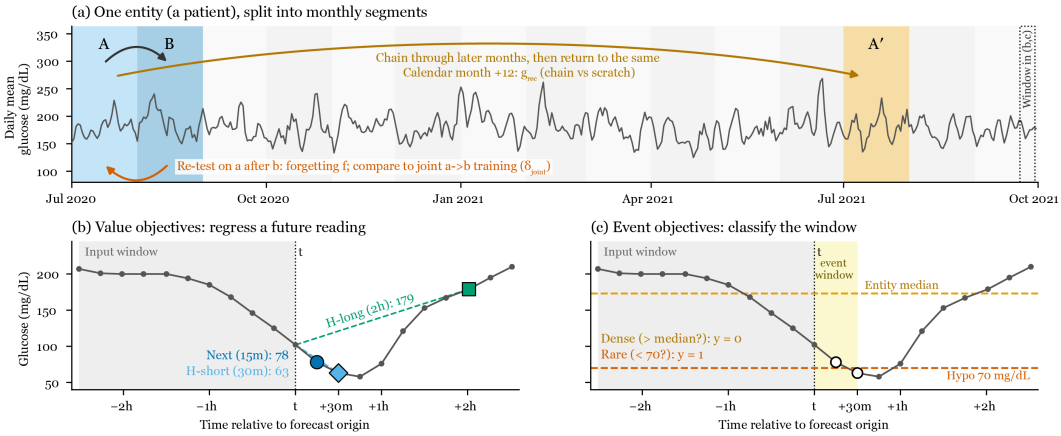


Figure 1: Protocol and objectives. (a) Each entity's stream is split into monthly temporal segments; models are fine-tuned sequentially (A→B), re-evaluated on A (forgetting, compared with joint A∪B training), and evaluated on the same calendar month one year later (recurrence gap). (b) Value objectives regress a future reading; (c) Event objectives classify whether any reading in the event window crosses a threshold. Shown for one glucose patient.

4 METHODS

4.1 DATASETS AND OBJECTIVES

Glucose (Granada T1D) contains 736 patients and 22.7M glucose readings sampled at ∼15-minute intervals (median calendar span of 334 days) (Rodriguez-Leon et al., 2023). Electricity load (UCI ElectricityLoadDiagrams) comprises 336 clients and 38.2M electricity meter readings sampled at ∼15-minute intervals (median of 1097 days with measurements) (Trindade, 2015). We construct WorldWeatherCities from Open-Meteo Historical Weather (ERA5-Land) with 345 cities and 45.4M temperature readings sampled at hourly intervals (median of 5479 days) (Zippenfenig, 2024; Muñoz-Sabater et al., 2021; Hersbach et al.).

We treat each patient, electricity meter, or city as an *entity* (one continuous univariate stream) and apply the same protocol per entity. We evaluate five objectives on each data stream: (1) *next-value* regression (MSE loss); (2) *short-horizon* value regression (glucose/electricity: 30 min; weather: 1 h; MSE); (3) *long-horizon* value regression (glucose: 2 h; electricity: 6 h; weather: 24 h; MSE); (4) *dense-event* classification (above-median within horizon: glucose/electricity 30 min, weather 6 h; BCE (primary)/AUROC (secondary)); (5) *rare-event* classification (hypoglycemia/peak load/extreme temperature within the same horizons; BCE (primary)/AUROC (secondary)). We use

non-overlapping 30-day segments starting from each entity's first reading (pairwise cohort) and calendar-month segments for sequential chains; entities need at least 24 months of data, and segments must pass domain-specific density gates (min. readings per segment: glucose 100, electricity 80, weather 48). For each eligible entity we sample up to 20 directed transitions $A \to B$ ($A < B$) uniformly at random. Sequential chains use the first n=35 entities passing stricter continuity filters (span, coverage, and readings per day).

4.2 Models and Training

We use five sequence learners {LSTM, GRU, RNN, CNN, Transformer} (default hidden size 128; input window L=24). Value targets are expressed in the same segment-normalized units as the inputs, so level shifts between segments do not reach the target. Event labels instead use *entity-level* thresholds fixed in raw units per entity (the entity's median for dense events; clinical 70 mg/dL for hypoglycemia; the entity's $q_{0.95}$ for peak load; its $q_{0.05}$/$q_{0.95}$ for temperature extremes), mapped into each segment's normalized units. Event base rates therefore vary with the segment's level, whereas value targets are segment-relative. We call this the objective's *reference* (leveraged in P5). We optimize with Adam (10^{-3}), batch size 64, $\leq$ 200 epochs, early stopping (patience 15), and ReduceLROnPlateau (factor 0.5, patience 5), with MSE for regression and BCE for event objectives. BCE reflects both ranking and calibration, AUROC ranking only; rising BCE with stable AUROC indicates forgotten calibration (including the label prior) rather than discrimination. In transfer steps (A$\to$B or chain transitions), model weights are continued while optimizer state is reinitialized.

4.3 Metrics

Primary evaluations maximize score, $\text{score} = -\mathcal{L}$, where $\mathcal{L}$ is MSE for regression objectives and BCE-with-logits for event objectives. Let $\text{score}_{A,B}$ be the score on held-out data from segment A after training through segment B ($A \leq B$). Absolute forgetting, F, is

$$F_{A\to B} = \text{score}_{A,A} - \text{score}_{A,B},$$

such that $F > 0$ means performance on A worsened after learning B (related to backward transfer (BWT) (Lopez-Paz & Ranzato, 2017)). Relative forgetting (F_{rel}) and the forgetting rate are

$$F_{\text{rel},A\to B} = \frac{F_{A\to B}}{|\text{score}_{A,A}| + \epsilon}, \qquad P(F > 0) = \frac{|\{A \to B : F_{A\to B} > 0\}|}{|\{A \to B\}|}$$

Pre-transfer skill compares segment-A loss to a simple baseline (persistence for regression; class-prior constant for classification) (Murphy, 1988; McFadden; Roulston & Smith, 2002):

$$\text{skill}_A = 1 - \frac{\mathcal{L}_{A,A}}{\mathcal{L}_{\text{baseline}}}.$$

Pairwise analyses retain transitions with $\text{skill}_A \geq 0$. After a full within-entity chain, the *savings ratio* is $(1 - T_{\text{relearn}}/T_{\text{scratch}})$, where T is epochs to reach within 5% of the scratch best validation loss on the first segment (else the full run length). To isolate the contribution of label-prior shift to event forgetting, we compute the forgetting of a predictor that knows only the base rate,

$$F^{\text{prior}}_{\text{rel},A\to B} = \frac{\text{BCE}(y_A; \bar{p}_B) - \text{BCE}(y_A; \bar{p}_A)}{\text{BCE}(y_A; \bar{p}_A)},$$

where $\bar{p}_m$ is the training-split positive rate of segment m and y_A are the test labels of A. We report the Spearman correlation between $F^{\text{prior}}_{\text{rel}}$ and observed F_{rel} across transitions, both raw and controlling for input shift (Wasserstein distance between A and B) (Mallasto & Feragen, 2017).

For failure-mode diagnostics we additionally compare relative forgetting under sequential $A \to B$ fine-tuning versus joint training on $A \cup B$, $\Delta_{\text{joint}} = F^{\text{seq}}_{\text{rel}} - F^{\text{joint}}_{\text{rel}}$, (positive $\to$ preventable interference), and a *recurrence gap*:

$$G_{\text{rec}} = \mathcal{L}^{\text{chain}}_r - \mathcal{L}^{\text{scratch}}_r,$$

the test loss on the first later segment r sharing the calendar month of $\mathcal{S}_1$ ($\geq$12 months later), for a model chained through $\mathcal{S}_1, \ldots, \mathcal{S}_{r-1}$ without training on r, minus a model trained on r from scratch. To separate prior drift from lost competence, we refit only an output offset b (logit $+ b$) or

Table 1: Median within-entity relative forgetting F_{rel} by objective (LSTM; n = pairwise transitions), with median chain savings $1 - T_{\text{relearn}}/T_{\text{scratch}}$ below. H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather). All architectures in Fig. 2 and Appendix 14.

OBJECTIVE	GLUCOSE (F_{rel} (n))	ELECTRICITY (F_{rel} (n))	WEATHER (F_{rel} (n))	SAVINGS (G/E/W)
Next-Value	0.02 (2791)	0.13 (5525)	0.00 (3309)	0.95 / 0.96 / 0.97
H-short (30m/30m/1h)	0.01 (2702)	0.22 (5539)	0.02 (4450)	0.94 / 0.94 / 0.97
H-long (2h/6h/24h)	0.01 (2717)	0.24 (5960)	0.09 (2754)	0.67 / 0.75 / 0.91
Dense event	0.14 (3021)	0.86 (5960)	1.92 (4716)	0.94 / 0.91 / 0.92
Rare event	0.09 (2753)	0.91 (4376)	1.41 (4666)	0.95 / 0.73 / 0.58

an affine map $a \cdot \text{logit} + b$ of the chained model on r's training split and scored on its test split; we report the fraction of the median gap closed, $1 - \text{med}\, G_{\text{rec}}^{\text{cal}} / \text{med}\, G_{\text{rec}}$. Positive G_{rec} means the chain underperforms a new fit on the returning segment. Unless noted, we report median F_{rel} within entity over transitions, and apply QC filters that drop saturated or unstable cells before typology summaries.

4.4 Evaluation Protocols

(P1) Pairwise forgetting: train on A, test on A, fine-tune on B ($A < B$), re-test on segment A (within-entity). *(P2) Sequential chains:* train across ordered monthly segments (n=35) and record $\text{score}_{A,B}$ trajectories and savings. *(P3) Failure-mode diagnostics:* Joint $A \cup B$ vs. sequential $A \to B$ (preventable interference); recurrence gap (return to the same calendar month ≥ 12 months later, with output calibration vs. scratch); CKA/probes as supportive representation checks. *(P4) Capacity and architecture ablations:* compare learner families and hidden size on fixed transitions. *(P5) Reference swap:* on the same pairwise transitions and seeds, (i) relabel dense events against each segment's own training-split median (segment reference; base rate approximately fixed), and (ii) train next-value models on entity-level rather than segment-level normalization (entity reference; level shifts reach the target). Pre-transfer skill is matched across conditions (LSTM; ≈ 400 transitions per domain). Pre-transfer skill is comparable across conditions (Appendix Table 19).

5 Forgetting is Objective-Dependent (Q1)

Under the same streams and Domain-IL protocol, median F_{rel} is small for value objectives (next, H-short, H-long) and large for event objectives (dense, rare). Table 1 reports within-entity forgetting by objective for the LSTM, and Figure 2 extends it to all five architectures. Architecture changes how much a model forgets, not which objectives it forgets: in every one of the 15 architecture $\times$ domain combinations, the smallest event-objective F_{rel} exceeds the largest value-objective F_{rel} (Appendix Table 14). Increasing capacity does not close the gap; rare-event forgetting grows with model size (Appendix Table 6). Below we rule out rarity, forecast difficulty, and the segment protocol as explanations (§5.1), and trace the gap to what each target is measured against (§5.2).

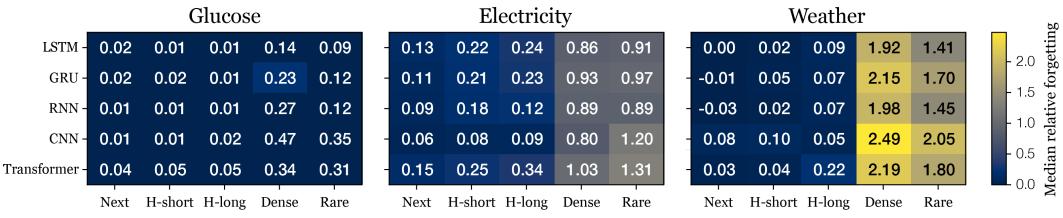


Figure 2: Median within-entity F_{rel} by objective across five architectures. Architecture changes magnitude but not the ordering of objectives.

5.1 CONTROLS

Rarity. Dense and rare forgetting are comparable in every domain (Table 1), so rarity is imprecise. The rarest event with a stable base rate, glucose hypoglycemia (7% of windows), forgets least of the six event cells, while weather dense events, whose base rate swings from 1% in January to nearly 100% in summer, forget most (Appendix Table 18).

Horizon. If forgetting followed forecast difficulty, long-horizon regression would approach event levels; instead H-long forgetting stays within 0.07 of H-short in every domain (Table 1; ladders in Appendix Figure 4, Table 12).

Protocol. Monthly segment blocks could exaggerate forgetting relative to online learning. Under an online sliding-window protocol, held-out performance usually improves rather than degrades; electricity peak events are the exception, with 20–32% degradation (weather values are saturated; Appendix Table 13, Figure 6).

5.2 MECHANISM: EVENT FORGETTING FOLLOWS THE MOVING REFERENCE

Value targets are expressed in segment-normalized units, whereas event labels use thresholds fixed per entity, so a shift in a segment's level changes the event base rate but not the value target (§4.3). Three results indicate that this difference in reference, rather than rarity or difficulty, drives the gap; predominantly, the shift in base rate can be measured, though a fixed threshold can also move the labeling rule without changing the base rate (§6, Case C).

Ranking survives, but calibration does not. After fine-tuning on B, event models rank A's windows almost as well as before ($|\Delta\mathrm{AUROC}| < 0.04$ for all learners except the transformer on electricity peaks and weather extremes), while BCE rises by up to $4.8\times$ (Appendix Table 5). What is lost is mainly the mapping from features to calibrated probabilities on A, not the ability to rank A's windows: $|\Delta\mathrm{AUROC}| < 0.04$ while BCE rises sharply (Appendix Table 5).

Forgetting tracks base-rate shift. Across all valid within-entity transitions, the forgetting of a predictor that knows only the base rate, $F_{\mathrm{rel}}^{\mathrm{prior}}$, ranks observed event forgetting with Spearman $\rho = 0.62$–0.87 in electricity and weather, nearly unchanged after controlling for the input-distribution distance between A and B ($\rho = 0.58$–0.84) and consistent across all five architectures (spread ≤ 0.07; Appendix Table 17). Glucose, whose event base rates barely move, shows the weakest dependence ($\rho = 0.35$–0.46). The base-rate-only predictor accounts for the ordering of forgetting rather than its magnitude: trained models forget several times more than it does, consistent with confident predictions paying more for a stale prior than a constant does.

Swapping the reference moves forgetting. We then manipulate the reference directly on the same transitions and seeds (P5; Appendix Table 19). Relabeling events against each segment's own quantiles reduces dense-event forgetting from 2.31 to 0.06 (weather) and from 0.93 to 0.16 (electricity), and removes it in glucose ($0.13 \rightarrow -0.05$). The same swap on rare events cuts peak forgetting from 1.13 to 0.13 and extreme forgetting from 1.35 to 0.12; skill is matched except for weather extremes, where segment-relative targets are harder, but the drop survives a skill-matched subset ($1.43 \rightarrow 0.06$; Appendix Table 19).

The event-objective penalty is therefore a property of how the objective is defined, not of events as such. The practical cost is real: alarm thresholds act on calibrated probabilities, and the low rare-event savings (electricity 0.73, weather 0.58; Table 1) fall exactly where base rates swing seasonally. Section 6 shows that the same label-prior drift underlies the recurrence gap, which output recalibration largely closes.

6 FORGETTING IS NOT HOMOGENEOUS (Q2)

Two diagnostics separate the ways sequential fine-tuning fails. Preventable interference Δ_{joint} asks whether joint training on A∪B would have avoided the damage. The recurrence gap G_{rec} asks whether a chained model underperforms a fresh fit when the same calendar month returns a year

	Next	H-short	H-long	Dense	Rare
Glucose	**L** 0.00 Δ +0.02 G 0.00	**L** +0.01 Δ +0.03 G 0.00	**L** +0.03 Δ +0.04 G 0.00	**I** +0.45 Δ +0.36 G +0.02	**L** +0.11‡ Δ +0.15 G 0.00
Electricity	**I** +0.19 Δ +0.21 G 0.00	**I** +0.24 Δ +0.26 G 0.00	**L** +0.12‡ Δ +0.17 G +0.01	**I** +0.55 Δ +0.51 G +0.02	**I** +0.43* Δ +0.35 G +0.01
Weather	**L** −0.35† Δ +0.06 G −0.01	**L** −0.39† Δ +0.08 G 0.00	**L** −0.11† Δ +0.01 G −0.01	**M** +0.35* Δ +0.34 G −0.02	**R** +0.01 Δ +0.13 G +0.34

Table 2: Failure-mode typology. Each cell: class (**L** low forgetting, **I** interference, **R** recurrence gap, **M** mixed) and median sequential F_{rel} across five learners; second line: preventable interference Δ_{joint} and recurrence gap G_{rec}. †negative forgetting (data-limited single-segment fits); ‡within 0.02 of the low-forgetting cutoff; *some learners computed after excluding saturated entities (e.g., 51–62% of cities were excluded in weather-dense). H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather).

later. Classifying each domain × objective cell on these diagnostics (Fig. 3; Table 2) yields two distinct signatures, which call for different remedies.

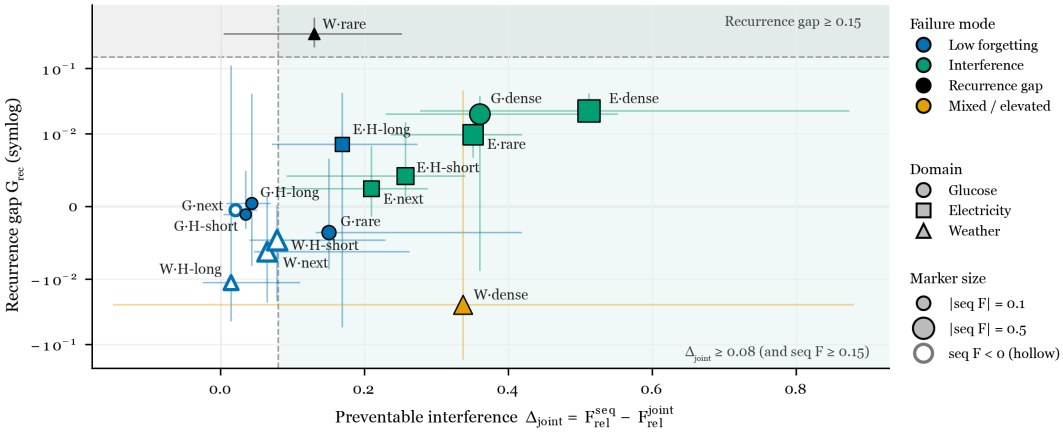


Figure 3: Typology diagnostic map. Each point is a domain × objective cell (median across five learners; whiskers span learners), placed by preventable interference Δ_{joint} and recurrence gap G_{rec}. Dashed lines mark the classification thresholds; hollow markers indicate negative forgetting.

Case study A: Electricity is interference-dominated. For next-value, H-short, dense, and rare objectives, sequential fine-tuning damages A ($F_{\text{rel}}^{\text{seq}} = 0.19$–$0.55$), joint training prevents most of it (helps 84–100% of entities), and $G_{\text{rec}} \leq 0.02$. Value damage is consistent with parameter overwrite. Event damage follows the entity-fixed reference. Peak forgetting ranks with the base-rate shift ($\rho = 0.85$), and segment-relative thresholds cut dense forgetting $0.93 \to 0.16$ and rare $1.13 \to 0.13$ (Appendix Tables 17, 19). H-long sits just below the low-forgetting cutoff.

Case study B: Weather seasonal tail events show a recurrence gap. Transfer between matching Januarys barely forgets ($F_{\text{rel}}^{\text{seq}} = 0.01$), yet a chain through the year underperforms scratch on the return ($G_{\text{rec}} = 0.34$; 0.22–0.58; Appendix Table 11). The gap tracks the prior. After near-zero extremes in spring/autumn, the chain predicts mean probability 0.12 on a January where extremes are common again. A logit offset closes 72% of the gap (69% with one week of data; Platt 77%); a residual ≈ 0.1 remains for the RNN, CNN, and transformer (Appendix Table 20). Other weather cells differ, however, as value objectives show negative forgetting and dense events are mixed and often saturated, so we treat that cell cautiously.

Case study C: Glucose is interference-dominated only for dense events. Dense events forget substantially (median 0.45) and joint training prevents it (94–100% of patients). Unlike electricity,

this tracks input shift more than the marginal prior (partial $\rho = 0.20$ vs. Wasserstein $\rho = 0.60$), yet disappears under a segment-relative threshold ($0.13 \rightarrow -0.05$): with segment-normalized inputs, a fixed raw threshold moves the labelling rule itself (Appendix Tables 17, 19). Value objectives stay near zero ($|F_{\mathrm{rel}}^{\mathrm{seq}}| \leq 0.03$); hypoglycemia is modest (0.11), except CNN/transformer (0.29/0.49).

The two signatures thus separate at least three mechanisms, each with a different remedy. Representational interference (electricity value objectives) is preventable with access to both segments. Reference-induced interference (event objectives with entity-fixed thresholds) disappears under segment-relative targets. Prior recurrence (weather extremes) is largely repaired by recalibrating when the regime returns. The standard forgetting metric detects only the first two: the recurrence gap is invisible to F_{rel}.

7 Forgetting can be recoverable (Q3)

Here we investigate whether forgetting can be recovered, and where in the model the damage sits. The answer depends on architecture: for the LSTM, event forgetting is concentrated in the readout and the label prior and is cheaply reversed, whereas for the transformer, and partly the CNN, damage extends into the representation.

Savings. After a full chain, relearning the first segment is fast for recurrent learners: LSTM savings are 0.91–0.97 for value and dense objectives in every domain (Table 1). Savings are lower for rare events in the seasonal domains (electricity 0.73, weather 0.58), exactly where base rates swing most (§5.2). Architecture changes this picture qualitatively, not only in magnitude. The transformer has negative rare-event savings in all three domains (-1.00 glucose, -2.44 electricity, -4.33 weather), and the CNN on electricity events (-0.40 rare, -0.33 dense). For these learners, a chained initialization relearns more slowly than a random one (Appendix Table 15). Savings measure recovery under full re-exposure to the forgotten segment; they do not show that competence was retained.

Recovery without re-exposure. A stricter test asks how much can be restored without retraining on old data. For the weather recurrence gap, refitting only an output offset on about one week of the returning month closes 69% of the gap, and 72% with the full training split (Appendix Table 20). Closure is near-complete for the LSTM and GRU (92–115%) but partial for the RNN, CNN, and transformer (44–64%), which retain a gap of $\approx$0.1 after recalibration.

Representation stability. Linear CKA before vs. after the chain drops more for rare-event than next-value objectives in 14 of 15 learner–domain cells, most for the transformer (Appendix Figure 8, Table 7).

Where the damage lives. On weather extremes (35 cities), we localize the damage by comparing the chained model with its own pre-chain state on the first segment (Appendix Table 21). For the LSTM, the representation still separates events. A linear probe on frozen post-chain features is as accurate as before (AUROC $0.92 \rightarrow 0.94$), retraining only the output head recovers all of the lost loss (median 102%), and even transplanting the original head back recovers 61%. The damage is in the readout. For the transformer, the representation itself has moved. Probe AUROC drops ($0.91 \rightarrow 0.85$), the original head no longer fits the new features (transplant recovers $\approx$0%), and head retraining recovers only 85% (55% for the parameter-matched variant). Interventions that protect the representation reduce forgetting for the full-size transformer (post-chain BCE $0.83 \rightarrow 0.63$ with either an encoder freeze or tiny replay; freezing also helps the parameter-matched variant, $0.65 \rightarrow 0.55$) but not for the LSTM (freezing: $0.68 \rightarrow 0.81$; replay: $0.68 \rightarrow 0.66$).

8 Implications

Forgetting on these temporal streams is objective-dependent: event objectives forget far more than value objectives under a matched protocol, and the gap follows how the target is referenced rather than rarity, horizon, or architecture (Q1). Diagnostics then split event cells into at least two signatures, preventable interference and a recurrence gap invisible to single-transition F_{rel} (Q2). Recoverability depends on where the damage sits. For recurrent learners it is often concentrated in the readout and label prior, but for the transformer it reaches the representation (Q3).

The practical consequence is that next-step regression (or any single default time-series loss) is a poor proxy for whether a temporal model "forgot." Before mitigating, ask which objective produced the drop and which failure signature it matches. Joint training diagnoses preventable interference; recalibration tests prior recurrence; freeze and replay probe whether the encoder is implicated. Choosing an event objective with entity-fixed thresholds knowingly buys operational relevance and, on these corpora, higher continual-learning cost.

A useful theory of forgetting on streams would need more than input-distribution distance. It should model (i) the geometry of the label process induced by the objective's reference, (ii) recurrence of regimes in calendar time, and (iii) how sequential updates interfere with earlier solutions. Shared temporal structure may still bound how bad forgetting can get for some objectives, but this is an informal capacity-of-shared-structure intuition rather than a measured "forgetting ceiling."

9 LIMITATIONS

Our claims are behavioral first. Mechanistic readings (head transplant, probes, freeze, replay) cover one domain and objective (weather extremes) and two architectures. Savings inherit early-stopping granularity and recalibration is cheap, but not label-free.

The three corpora are matched operationally, one univariate stream per patient, meter, or city, but not semantically. CGM tracks a single physiology, meters aggregate household demand, and city temperatures are environmental fields. These differences likely shape *which* signature appears even under a shared Domain-IL protocol. We treat cross-corpus heterogeneity as evidence that failure modes are not monolithic, not as a taxonomy of domains. Disentangling objective structure from the stream-generating process would need matched objectives on streams with controlled aggregation, which we leave to future work.

Additional limits. We study univariate structured streams, not multimodal EHR or foundation-model continual learning. Monthly segments are not fully online streaming; an online control usually improves rather than degrades, with electricity peaks as the main exception. Event evaluation inherits thresholding. Ranking (AUROC) and detection scores can dissociate evaluation, and BCE-only forgetting can overstate operational failure under imbalance. Weather dense and ultra-rare cells can saturate, so we applied QC filters. Transition-level intervals treat resampled pairs as independent and can be anti-conservative so entity-level inference is safer. Transformer savings are protocol-sensitive. Absolute forgetting is smaller on glucose than on the seasonal domains, so severity is domain-dependent even when ordering by objective is not.

10 CONCLUSION

Not all forgetting is equal. On structured temporal streams, catastrophic forgetting is conditioned on the prediction objective. Event targets with entity-fixed references forget far more than segment-relative value targets, for reasons that a reference swap can move. Within event forgetting, diagnostics separate preventable interference from a recurrence gap that single-transition metrics miss. The loss is often recoverable and it is cheap when damage sits in the readout and prior, but less so when it reaches the representation. Three corpora suffice to show that these signatures arise under a matched protocol, but they do not yield a complete domain taxonomy. Continual learning on streams should therefore be evaluated as objective-conditioned and diagnostically categorized, not as a uniform property of sequential data.

AI USE STATEMENT

Initial literature review, ideation, and concept development were conducted without the use of LLMs. Generative AI tools were subsequently used to explore experimental opportunities, assist with software development, and improve the clarity and conciseness of the manuscript. All LLM-generated code was independently reviewed, verified, and tested for correctness by the lead author.

AI-assisted literature-search and synthesis tools were used to support the literature review; however, all cited sources were independently reviewed before inclusion, and all resulting claims were verified against the original literature. All definitions, ideas, and scientific insights were developed by the

authors before AI assistance and were reviewed and revised by the authors assistance. The authors take full responsibility for the final text, claims, analyses, code, and other artifacts presented in this work.

Reproducibility statement

All datasets used in this study are publicly available. As the weather dataset is from multiple locations, we provide a script for collecting and building the weather dataset. The code used for the experiments will be made publicly available.

Beyond access to all the experimental materials, reproducibility may be impacted by experiment sampling (entities, pairs, seeds), however we can release checksums of raw electricity and assembled weather. We will provide a list of the client/city information and a list of the 35 chain entities per domain. The glucose dataset is available upon request from https://zenodo.org/records/10713570.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Kyle Aitken, Marina Garrett, Shawn Olsen, and Stefan Mihalas. The geometry of representational drift in natural and artificial neural networks. *PLOS Computational Biology*, 18(11):e1010716, 2022.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.

Akash Choudhuri, Hankyu Jang, Alberto M. Segre, Philip M. Polgreen, Kishlay Jha, and Bijaya Adhikari. Continually-Adaptive Representation Learning Framework for Time-Sensitive Healthcare Applications. *Proceedings of the ... ACM International Conference on Information & Knowledge Management. ACM International Conference on Information and Knowledge Management*, 2023:4538–4544, 2023. ISSN 2155-0751. doi: 10.1145/3583780.3615464. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11920885/.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing Representation Forgetting in Supervised and Unsupervised Continual Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16691–16700, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01621. URL https://ieeexplore.ieee.org/document/9879096/.

Wenzhang Du. Mitigating Catastrophic Forgetting in Streaming Generative and Predictive Learning via Stateful Replay, November 2025. URL http://arxiv.org/abs/2511.17936. arXiv:2511.17936 [cs].

Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni. Don't forget, there is more than forgetting: new metrics for Continual Learning, October 2018. URL http://arxiv.org/abs/1810.13166. arXiv:1810.13166 [cs].

Hermann Ebbinghaus (1885). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences*, 20(4):155–156, October 2013. ISSN 0972-7531. doi: 10.5214/ans.0972.7531.200408. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4117135/.

Craig Fellers and Benjamin C Storm. The saving enhanced memory effect can be observed when only a subset of items are saved. *Memory & Cognition*, 52(6):1325–1337, 2024.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks, March 2015. URL http://arxiv.org/abs/1312.6211. arXiv:1312.6211 [stat].

Liangxuan Guo, Yang Chen, and Shan Yu. Out-of-distribution forgetting: vulnerability of continual learning to intra-class distribution shift. In *International Conference on Pattern Recognition*, pp. 111–127. Springer, 2024.

James D. Hamilton. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384, 1989. ISSN 0012-9682. doi: 10.2307/1912559. URL https://www.jstor.org/stable/1912559.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. doi: 10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3803.

Steven C.H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, October 2021. ISSN 09252312. doi: 10.1016/j.neucom.2021.04.112. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231221006706.

Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae, and Sungroh Yoon. New insights for the stability-plasticity dilemma in online continual learning. *arXiv preprint arXiv:2302.08741*, 2023.

Hankyul Kang, Gregor Seifer, Donghyun Lee, and Jongbin Ryu. Do your best and get enough rest for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10077–10086, 2025.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring Catastrophic Forgetting in Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11651. URL https://ojs.aaai.org/index.php/AAAI/article/view/11651.

Pouya Khodaee, Herna L Viktor, and Wojtek Michalowski. Knowledge transfer in lifelong machine learning: a systematic literature review. *Artificial Intelligence Review*, 57(8):217, 2024.

Chang-Jin Kim and Daniel C. R. Halbert. *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*. URL https://direct.mit.edu/books/monograph/3265/State-Space-Models-with-Regime-SwitchingClassical.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114. URL http://arxiv.org/abs/1612.00796. arXiv:1612.00796 [cs].

Boris Kovatchev, Marc Breton, and William Clarke. Analytical methods for the retrieval and interpretation of continuous glucose monitoring data in diabetes. *Methods in enzymology*, 454:69–86, 2009.

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization. *arXiv preprint arXiv:2308.11958*, 2023.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3057446. URL http://arxiv.org/abs/1909.08383. arXiv:1909.08383 [cs].

Chiara Lanza, Roberto Pereira, Marco Miozzo, Eduard Angelats, and Paolo Dini. Degradation of feature space in continual learning. *arXiv preprint arXiv:2602.06586*, 2026.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual Learning in the Teacher-Student Setup: Impact of Task Similarity, July 2021. URL http://arxiv.org/abs/2107.04384. arXiv:2107.04384 [stat.ML].

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on Forgetting and Generalization of Continual Learning, February 2023. URL http://arxiv.org/abs/2302.05836. arXiv:2302.05836 [cs.LG].

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Aojun Lu, Hangjie Yuan, Tao Feng, and Yanan Sun. Rethinking the stability-plasticity trade-off in continual learning from an architectural perspective. *arXiv preprint arXiv:2506.03951*, 2025.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online Continual Learning in Image Classification: An Empirical Survey, October 2021. URL http://arxiv.org/abs/2101.10423. arXiv:2101.10423 [cs.LG].

Andrey Malinin, Neil Band, Ganshin, Alexander, German Chesnokov, Yarin Gal, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Roginskiy, Denis, Mariya Shmatova, Panos Tigas, and Boris Yangel. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks, February 2022. URL http://arxiv.org/abs/2107.07455. arXiv:2107.07455 [cs.LG].

Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/7a006957be65e608e863301eb98e1808-Abstract.html.

Ignacio Cabrera Martin, Subhaditya Mukherjee, Almas Baimagambetov, Joaquin Vanschoren, and Nikolaos Polatidis. Evolving Machine Learning in Non-Stationary Environments: A Unified Survey of Drift, Forgetting, and Adaptation, January 2026. URL http://arxiv.org/abs/2505.17902. arXiv:2505.17902 [cs] version: 3.

Daniel McFadden. Conditional Logit Analysis of Qualitative Choice Behavior.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Timothy Nguyen, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Architecture Matters in Continual Learning, February 2022. URL http://arxiv.org/abs/2202.00275. arXiv:2202.00275 [cs].

Allan H. Murphy. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. December 1988. ISSN 1520-0493. URL https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, Brecht Martens, Diego G. Miralles, María Piles, Nemesio J. Rodríguez-Fernández, Ervin Zsoter, Carlo Buontempo, and Jean-Noël Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, September 2021. ISSN 1866-3508. doi: 10.5194/essd-13-4349-2021. URL https://essd.copernicus.org/articles/13/4349/2021/.

Anshul Nasery, Soumyadeep Thakur, Vihari Piratla, Abir De, and Sunita Sarawagi. Training for the Future: A Simple Gradient Interpolation Loss to Generalize Along Time. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19198–19209. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/a02ef8389f6d40f84b50504613117f88-Abstract.html.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, pp. 381–405. PMLR, October 2019. URL https://proceedings.mlr.press/v106/nestor19a.html.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. ISSN 1558-2256. doi: 10.1109/5.18626. URL https://ieeexplore.ieee.org/abstract/document/18626.

Ciro Rodriguez-Leon, Maria Dolores Aviles Perez, Oresti Banos, Miguel Quesada-Charneco, Pablo J Lopez-Ibarra, Claudia Villalonga, and Manuel Munoz-Torres. T1DiabetesGranada: a longitudinal multi-modal dataset of type 1 diabetes mellitus, October 2023. URL https://zenodo.org/records/10050944.

Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. June 2002. ISSN 1520-0493. URL https://journals.ametsoc.org/view/journals/mwre/130/6/1520-0493_2002_130_1653_epfuit_2.0.co_2.xml.

Shiori Sagawa, Pang Wei Koh, Fanny Yang, Hongseok Namkoong, Jiashi Feng, Kate Saenko, Percy Liang, Sarah Bird, and Sergey Levine. NeurIPS DistShift Workshop 2021. In *Distribution shifts: connecting methods and applications (DistShift)*. URL https://sites.google.com/view/distshift2021.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning, November 2019. URL http://arxiv.org/abs/1812.05159. arXiv:1812.05159 [cs].

Khoa Tran and Simon S Woo. Efficient unlearning through maximizing relearning convergence delay. *arXiv preprint arXiv:2604.09391*, 2026.

Artur Trindade. ElectricityLoadDiagrams20112014, 2015. URL https://archive.ics.uci.edu/dataset/321.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, December 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00568-3. URL https://www.nature.com/articles/s42256-022-00568-3.

Gido M Van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*, 2024.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application, February 2024. URL http://arxiv.org/abs/2302.00487. arXiv:2302.00487 [cs.LG].

Patrick Zippenfenig. Open-Meteo.com Weather API. December 2024. doi: 10.5281/zenodo.14582479. URL https://zenodo.org/records/14582479.

Table 3: The main gap (LSTM shown here; additional architectures can be found in the appendix). Shown are relative forgetting (F_{rel}), the number of transitions (n), and savings

OBJECTIVE	GLUCOSE (F_{rel} (n))	ELECTRICITY (F_{rel} (n))	WEATHER (F_{rel} (n))	SAVINGS (G/E/W)
Next-Value	0.02 (2791)	0.13 (5525)	0.00 (3309)	0.95 / 0.96 / 0.97
H-short (30m/30m/30m)	0.01 (2702)	0.22 (5539)	0.02 (4450)	0.94 / 0.94 / 0.97
H-long (2h/2h/24h)	0.01 (2717)	0.30 (5835)	0.09 (2754)	0.67 / 0.89 / 0.91
Dense event	0.14 (3021)	0.86 (5960)	1.92 (4716)	0.94 / 0.91 / 0.92
Rare event	0.09 (2753)	0.91 (4376)	1.41 (4666)	0.95 / 0.73 / 0.58

Table 4: Dense versus rare δs

DOMAIN	LEARNER	NEXT	DENSE EVENT	RARE EVENT	DENSE OVER RARE	RARE OVER NEXT
Glucose	LSTM	0.020	0.143	0.093	1.529	4.672
Glucose	GRU	0.021	0.230	0.119	1.921	5.671
Glucose	RNN	0.013	0.265	0.124	2.138	9.690
Glucose	CNN	0.013	0.472	0.351	1.345	27.176
Glucose	Transformer	0.043	0.336	0.308	1.090	7.154
Electricity	LSTM	0.131	0.857	0.908	0.943	6.952
Electricity	GRU	0.107	0.932	0.973	0.958	9.121
Electricity	RNN	0.087	0.893	0.893	1.000	9.958
Electricity	CNN	0.064	0.801	1.195	0.670	18.544
Electricity	Transformer	0.152	1.029	1.311	0.785	8.649
Weather	LSTM	0.002	1.924	1.406	1.368	634.088
Weather	GRU	-0.010	2.148	1.697	1.265	-172.829
Weather	RNN	-0.030	1.978	1.448	1.367	-47.925
Weather	CNN	0.078	2.488	2.053	1.212	26.393
Weather	Transformer	0.031	2.190	1.803	1.215	58.230

A Appendix

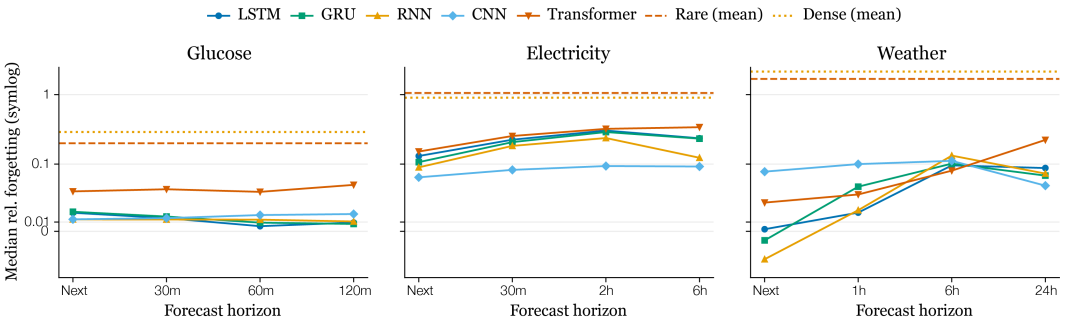


Figure 4: Horizon ladders: F_{rel} as the forecast horizon grows (shared symlog axis). Longer horizons do not produce event-level forgetting.

Fig/Table: online sliding stress

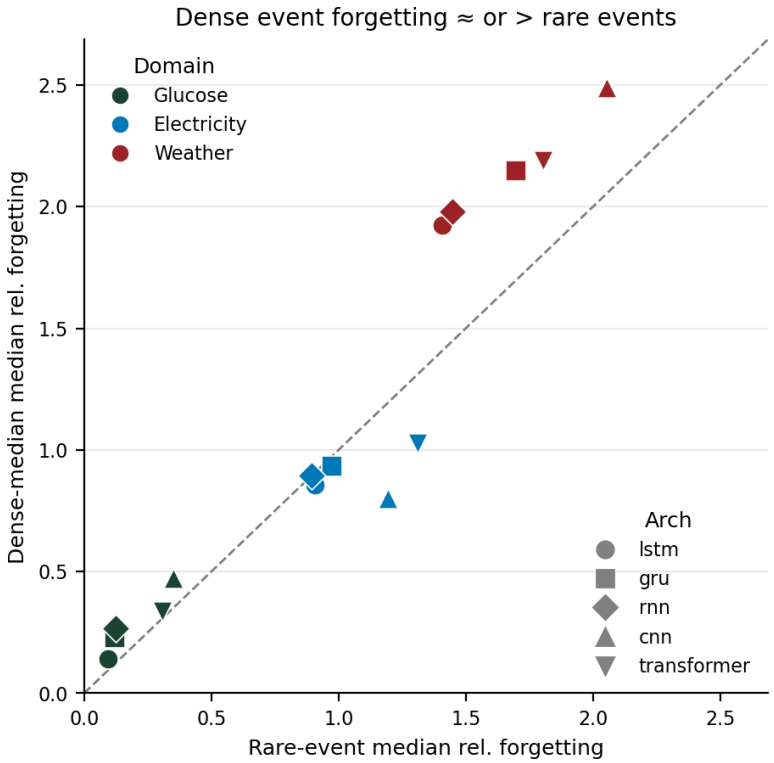


Figure 5: Within-entity relative forgetting by objective.

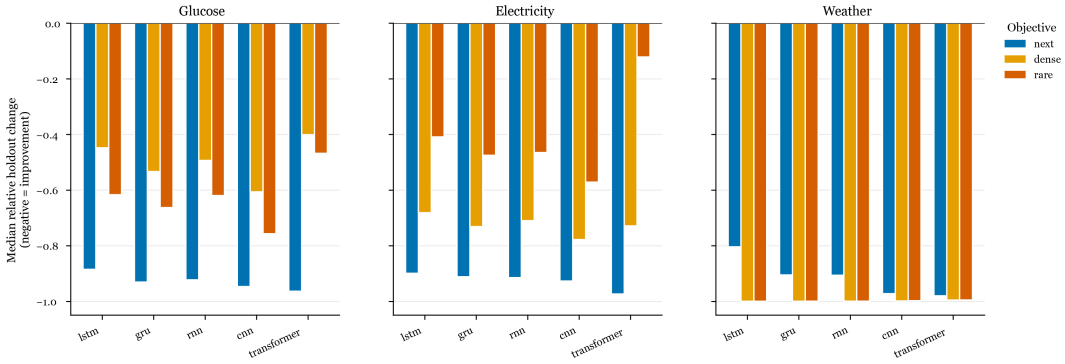


Figure 6: Online sliding-window stress (control): holdout usually improves.

Table 5: Discrimination vs. calibration after sequential fine-tuning ($n = 35$ entities). AUROC_0: AUROC on A before B; ΔAUROC: median drop after B; BCE F_{rel}: median relative BCE increase; BCE forgets: fraction of entities with BCE forgetting; ρ: correlation between AUROC and BCE changes. BCE forgetting is large while ranking (AUROC) is largely preserved ($|\Delta\text{AUROC}| < 0.04$), except the transformer on electricity peak and rare weather seasonal tail events.

Domain / task	Learner	AUROC_0	ΔAUROC	BCE F_{rel}	BCE forgets	ρ
Glucose dense	LSTM	0.96	−0.005	+0.29	69%	+0.44
	GRU	0.97	+0.001	+0.54	86%	+0.37
	RNN	0.97	+0.002	+0.42	80%	+0.40
	CNN	0.98	+0.005	+0.52	94%	+0.39
	Transformer	0.96	+0.004	+0.41	100%	+0.37
Glucose rare (hypo)	LSTM	0.96	−0.014	+0.14	59%	+0.28
	GRU	0.96	−0.008	+0.44	72%	+0.22
	RNN	0.96	−0.009	+0.41	53%	+0.18
	CNN	0.98	+0.001	+0.52	72%	+0.06
	Transformer	0.97	+0.007	+0.45	75%	+0.68
Electricity dense	LSTM	0.99	+0.005	+3.22	94%	+0.15
	GRU	0.99	+0.006	+4.43	94%	+0.27
	RNN	0.99	+0.005	+3.38	94%	+0.20
	CNN	0.99	+0.006	+4.15	100%	+0.14
	Transformer	0.99	+0.010	+4.03	97%	+0.44
Electricity rare (peak)	LSTM	0.92	+0.003	+4.58	97%	+0.35
	GRU	0.90	+0.011	+4.76	100%	+0.23
	RNN	0.92	−0.001	+4.15	90%	+0.09
	CNN	0.91	+0.039	+6.81	97%	+0.26
	Transformer	0.92	+0.195	+5.58	100%	+0.35
Weather dense	LSTM	0.98	−0.002	−0.40	24%	+0.16
	GRU	0.98	−0.002	−0.38	28%	−0.07
	RNN	0.98	−0.008	−0.27	36%	+0.13
	CNN	0.91	−0.017	+0.16	60%	+0.45
	Transformer	0.85	+0.020	+0.11	56%	+0.12
Weather rare (extreme)	LSTM	0.88	+0.013	+1.17	76%	+0.54
	GRU	0.90	−0.006	+1.19	88%	+0.57
	RNN	0.94	+0.033	+0.55	68%	+0.78
	CNN	0.94	+0.012	+3.94	97%	+0.30
	Transformer	0.93	+0.101	+1.77	91%	+0.39

Table 6: Capacity ablation: median relative forgetting by hidden size (columns). Weather was run at a subset of sizes. Increasing capacity does not close the next-vs-rare gap; rare-event forgetting grows with size.

Domain (objective)	Learner	4	8	16	32	64	128	256
Glucose (next)	LSTM	0.00	0.00	−0.01	0.00	0.00	0.00	0.00
	GRU	−0.01	0.00	0.00	0.00	0.00	−0.01	0.01
	RNN	−0.01	0.00	0.00	0.00	0.01	0.01	0.03
	CNN	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Transformer	0.02	0.02	0.02	0.03	0.03	0.06	0.06
Glucose (hypo)	LSTM	0.01	0.02	0.01	0.02	0.06	0.02	0.18
	GRU	0.04	0.05	0.05	0.11	0.02	0.05	0.13
	RNN	0.02	0.01	0.06	0.10	0.13	0.13	0.14
	CNN	0.07	0.12	0.15	0.20	0.24	0.24	0.31
	Transformer	0.12	0.18	0.22	0.23	0.31	0.42	0.46
Electricity (next)	LSTM	0.03	0.03	0.08	0.07	0.13	0.08	0.11
	GRU	0.01	0.03	0.04	0.05	0.06	0.12	0.15
	RNN	0.00	0.01	0.02	0.05	0.08	0.11	0.07
	CNN	0.00	0.01	0.01	0.02	0.03	0.04	0.05
	Transformer	0.06	0.06	0.11	0.10	0.12	0.12	0.13
Electricity (peak)	LSTM	0.49	0.43	0.63	0.69	1.03	0.64	1.28
	GRU	0.64	0.80	0.85	0.65	0.95	1.07	1.17
	RNN	0.59	0.62	1.05	0.65	0.70	0.87	0.81
	CNN	0.70	0.85	0.82	1.04	1.07	0.95	1.23
	Transformer	0.75	0.91	1.03	0.99	1.10	0.89	1.61
Weather (next)	LSTM	−0.15	–	–	−0.01	–	−0.03	0.01
	GRU	−0.19	–	–	−0.01	–	0.01	0.03
	RNN	−0.21	–	–	−0.05	–	−0.03	0.05
	CNN	−0.01	–	–	0.01	–	0.08	0.13
	Transformer	0.01	–	–	0.01	–	0.02	0.04
Weather (extreme)	LSTM	0.16	–	–	1.03	–	1.64	1.84
	GRU	0.39	–	–	0.67	–	1.47	1.57
	RNN	0.22	–	–	0.83	–	1.34	1.69
	CNN	0.50	–	–	1.17	–	1.75	2.00
	Transformer	0.52	–	–	1.00	–	1.49	1.44

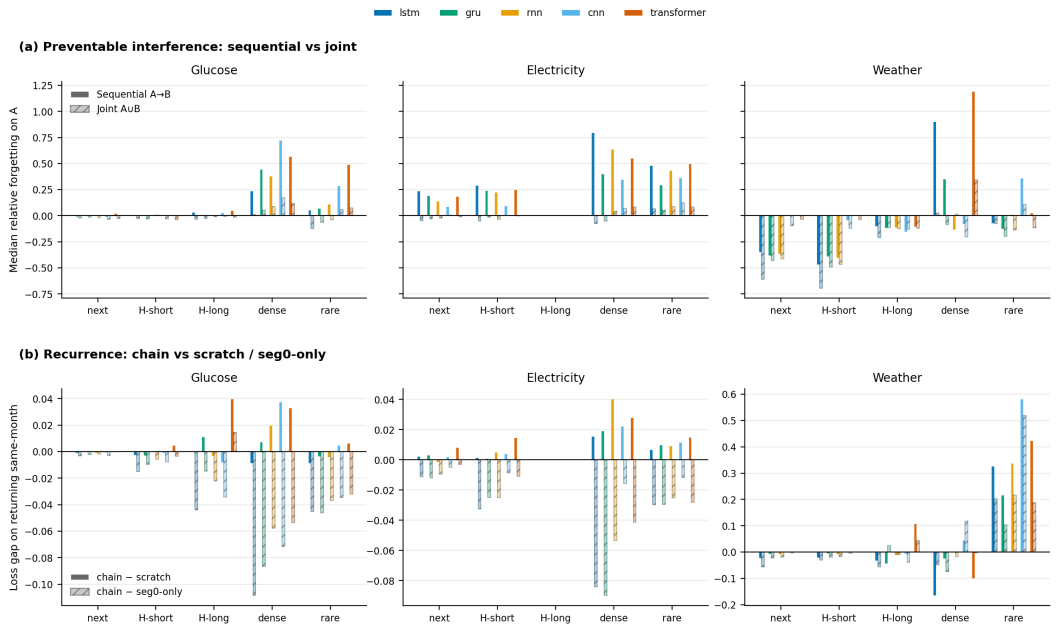


Figure 7: Panel (a) is sequential vs. joint forgetting and panel (b) is the recurrence gaps, with one shared learner legend.

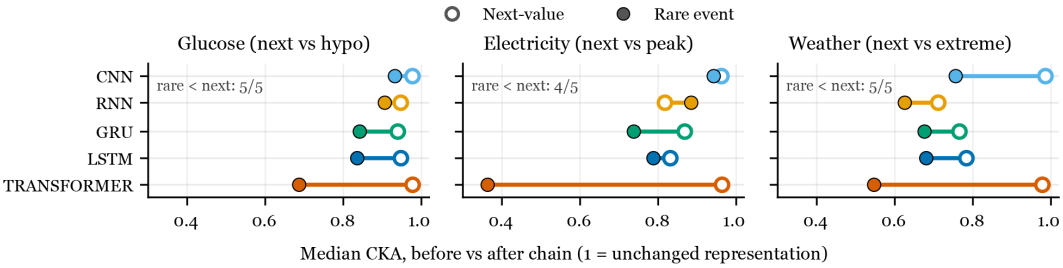


Figure 8: Representation stability under sequential fine-tuning. Median linear CKA between representations before and after the within-entity chain ($n = 35$ entities per domain), for next-value (hollow) and rare-event (filled) objectives. Rare-event objectives reduce representational similarity more than next-value in 14 of 15 learner–domain cells; the full-size transformer shows the largest drops. Parameter-matched transformer and per-objective bars in Appendix 7.

Table 7: Linear CKA between representations before and after the within-entity chain (median over $n = 35$ entities). Rare events: hypo (glucose), peak (electricity), extreme (weather). Δ = rare $-$ next. Next-value controls were not run for the parameter-matched transformer.

Domain	Learner	Params	CKA next	CKA rare	Δ	SD (rare)
Glucose	LSTM	67,201	0.95	0.84	-0.11	0.10
	GRU	50,433	0.94	0.84	-0.10	0.12
	RNN	16,897	0.95	0.91	-0.04	0.17
	CNN	25,089	0.98	0.93	-0.05	0.08
	Transformer	396,929	0.98	0.69	-0.29	0.22
	Transformer (param.-matched)	67,137	–	0.75	–	0.17
Electricity	LSTM	67,201	0.83	0.79	-0.04	0.19
	GRU	50,433	0.87	0.74	-0.13	0.16
	RNN	16,897	0.82	0.88	$+0.07$	0.19
	CNN	25,089	0.96	0.94	-0.02	0.22
	Transformer	396,929	0.96	0.36	-0.60	0.23
	Transformer (param.-matched)	67,137	–	0.56	–	0.23
Weather	LSTM	67,201	0.78	0.68	-0.10	0.18
	GRU	50,433	0.77	0.68	-0.09	0.20
	RNN	16,897	0.71	0.63	-0.09	0.21
	CNN	25,089	0.99	0.76	-0.23	0.17
	Transformer	396,929	0.98	0.55	-0.43	0.25
	Transformer (param.-matched)	67,137	–	0.68	–	0.23

Table 8: Dense vs. rare event forgetting for all architectures (median within-entity F_{rel}). Differences from next-value are reported instead of ratios, which are unstable when next-value forgetting is near zero.

Domain	Learner	Next	Dense	Rare	Dense$-$next	Rare$-$next
Glucose	LSTM	0.02	0.14	0.09	$+0.12$	$+0.07$
	GRU	0.02	0.23	0.12	$+0.21$	$+0.10$
	RNN	0.01	0.27	0.12	$+0.25$	$+0.11$
	CNN	0.01	0.47	0.35	$+0.46$	$+0.34$
	Transformer	0.04	0.34	0.31	$+0.29$	$+0.27$
Electricity	LSTM	0.13	0.86	0.91	$+0.73$	$+0.78$
	GRU	0.11	0.93	0.97	$+0.83$	$+0.87$
	RNN	0.09	0.89	0.89	$+0.80$	$+0.80$
	CNN	0.06	0.80	1.20	$+0.74$	$+1.13$
	Transformer	0.15	1.03	1.31	$+0.88$	$+1.16$
Weather	LSTM	0.00	1.92	1.41	$+1.92$	$+1.40$
	GRU	-0.01	2.15	1.70	$+2.16$	$+1.71$
	RNN	-0.03	1.98	1.45	$+2.01$	$+1.48$
	CNN	0.08	2.49	2.05	$+2.41$	$+1.98$
	Transformer	0.03	2.19	1.80	$+2.16$	$+1.77$

Table 9: Drift-vs-interference diagnostics per learner, electricity. $F_{\mathrm{rel}}^{\mathrm{seq}}$: forgetting on A after sequential A→B; $F_{\mathrm{rel}}^{\mathrm{joint}}$: after joint A∪B; $\Delta_{\mathrm{joint}} = F_{\mathrm{rel}}^{\mathrm{seq}} - F_{\mathrm{rel}}^{\mathrm{joint}}$; Joint helps: fraction of entities where joint training preserves A better. G_{rec}: loss gap on the same calendar month $\geq$12 months later, chain minus scratch (scr.) or chain minus the segment-0-only model (seg0). n: entities after excluding saturated entities (of 35). H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather).

Objective	Learner	$F_{\mathrm{rel}}^{\mathrm{seq}}$	$F_{\mathrm{rel}}^{\mathrm{joint}}$	Δ_{joint}	Joint helps	G_{rec} (scr.)	G_{rec} (seg0)	n
Next	LSTM	$+0.24$	-0.05	$+0.29$	100%	$+0.002$	-0.011	35
	GRU	$+0.20$	-0.03	$+0.23$	97%	$+0.003$	-0.012	35
	RNN	$+0.14$	-0.02	$+0.16$	100%	-0.001	-0.009	35
	CNN	$+0.09$	0.00	$+0.08$	91%	$+0.002$	-0.005	35
	Transformer	$+0.19$	-0.01	$+0.20$	94%	$+0.008$	-0.003	35
H-short	LSTM	$+0.29$	-0.05	$+0.34$	100%	$+0.001$	-0.033	35
	GRU	$+0.24$	-0.01	$+0.26$	97%	$+0.001$	-0.025	35
	RNN	$+0.23$	-0.04	$+0.26$	100%	$+0.005$	-0.025	35
	CNN	$+0.09$	0.00	$+0.09$	91%	$+0.004$	-0.009	35
	Transformer	$+0.25$	0.00	$+0.25$	91%	$+0.015$	-0.011	35
H-long	LSTM	$+0.16$	-0.10	$+0.26$	89%	-0.020	-0.167	35
	GRU	$+0.12$	-0.15	$+0.27$	100%	$+0.009$	-0.143	35
	RNN	$+0.10$	-0.05	$+0.15$	83%	-0.053	-0.148	35
	CNN	$+0.07$	-0.01	$+0.07$	89%	$+0.016$	-0.076	35
	Transformer	$+0.20$	0.00	$+0.20$	74%	$+0.041$	-0.082	35
Dense	LSTM	$+0.80$	-0.08	$+0.87$	100%	$+0.016$	-0.084	35
	GRU	$+0.40$	-0.05	$+0.45$	100%	$+0.019$	-0.090	35
	RNN	$+0.64$	$+0.04$	$+0.60$	94%	$+0.041$	-0.053	35
	CNN	$+0.35$	$+0.07$	$+0.28$	91%	$+0.022$	-0.016	35
	Transformer	$+0.55$	$+0.08$	$+0.47$	97%	$+0.028$	-0.041	35
Rare	LSTM	$+0.48$	$+0.07$	$+0.41$	84%	$+0.007$	-0.030	32/35
	GRU	$+0.29$	$+0.06$	$+0.24$	84%	$+0.010$	-0.030	32/35
	RNN	$+0.43$	$+0.09$	$+0.35$	84%	$+0.009$	-0.025	32/35
	CNN	$+0.37$	$+0.13$	$+0.24$	88%	$+0.012$	-0.012	32/35
	Transformer	$+0.50$	$+0.08$	$+0.42$	88%	$+0.015$	-0.028	32/35

Table 10: Drift-vs-interference diagnostics per learner, glucose. $F_{\text{rel}}^{\text{seq}}$: forgetting on A after sequential A→B; $F_{\text{rel}}^{\text{joint}}$: after joint A∪B; $\Delta_{\text{joint}} = F_{\text{rel}}^{\text{seq}} - F_{\text{rel}}^{\text{joint}}$; Joint helps: fraction of entities where joint training preserves A better. G_{rec}: loss gap on the same calendar month $\geq$12 months later, chain minus scratch (scr.) or chain minus the segment-0-only model (seg0). n: entities after excluding saturated entities (of 35). H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather).

Objective	Learner	$F_{\text{rel}}^{\text{seq}}$	$F_{\text{rel}}^{\text{joint}}$	Δ_{joint}	Joint helps	G_{rec} (scr.)	G_{rec} (seg0)	n
Next	LSTM	0.00	−0.02	+0.02	66%	−0.001	−0.003	35
	GRU	+0.01	−0.01	+0.02	69%	0.000	−0.002	35
	RNN	0.00	−0.02	+0.01	71%	−0.001	−0.002	35
	CNN	−0.01	−0.03	+0.02	74%	−0.001	−0.003	35
	Transformer	+0.02	−0.03	+0.05	89%	0.000	0.000	35
H-short	LSTM	0.00	−0.03	+0.03	63%	−0.002	−0.015	35
	GRU	+0.01	−0.03	+0.03	80%	−0.003	−0.010	35
	RNN	+0.01	0.00	+0.01	60%	0.000	−0.006	35
	CNN	+0.01	−0.03	+0.04	74%	−0.001	−0.008	35
	Transformer	0.00	−0.04	+0.04	77%	+0.005	−0.003	35
H-long	LSTM	+0.03	−0.03	+0.07	74%	0.000	−0.044	35
	GRU	+0.01	−0.03	+0.04	74%	+0.011	−0.015	35
	RNN	0.00	−0.01	+0.01	80%	−0.003	−0.022	35
	CNN	+0.03	−0.01	+0.04	60%	−0.008	−0.034	35
	Transformer	+0.05	−0.01	+0.07	86%	+0.040	+0.015	35
Dense	LSTM	+0.24	+0.01	+0.23	100%	−0.009	−0.109	35
	GRU	+0.45	+0.05	+0.39	94%	+0.007	−0.087	35
	RNN	+0.38	+0.09	+0.29	94%	+0.020	−0.058	35
	CNN	+0.72	+0.17	+0.55	97%	+0.037	−0.072	35
	Transformer	+0.57	+0.12	+0.45	94%	+0.033	−0.053	35
Rare	LSTM	+0.06	−0.12	+0.18	80%	−0.009	−0.045	35
	GRU	+0.07	−0.06	+0.13	89%	−0.004	−0.046	35
	RNN	+0.11	−0.04	+0.15	91%	−0.004	−0.037	35
	CNN	+0.29	+0.06	+0.23	86%	+0.005	−0.035	35
	Transformer	+0.49	+0.07	+0.42	80%	+0.006	−0.032	35

Table 11: Drift-vs-interference diagnostics per learner, weather. $F_{\text{rel}}^{\text{seq}}$: forgetting on A after sequential A→B; $F_{\text{rel}}^{\text{joint}}$: after joint A∪B; $\Delta_{\text{joint}} = F_{\text{rel}}^{\text{seq}} - F_{\text{rel}}^{\text{joint}}$; Joint helps: fraction of entities where joint training preserves A better. G_{rec}: loss gap on the same calendar month ≥12 months later, chain minus scratch (scr.) or chain minus the segment-0-only model (seg0). n: entities after excluding saturated entities (of 35). H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather).

Objective	Learner	$F_{\text{rel}}^{\text{seq}}$	$F_{\text{rel}}^{\text{joint}}$	Δ_{joint}	Joint helps	G_{rec} (scr.)	G_{rec} (seg0)	n
Next	LSTM	−0.35	−0.61	+0.26	86%	−0.022	−0.056	35
	GRU	−0.38	−0.43	+0.05	77%	−0.006	−0.023	35
	RNN	−0.36	−0.41	+0.05	80%	−0.006	−0.019	35
	CNN	0.00	−0.10	+0.09	86%	−0.002	−0.003	35
	Transformer	+0.01	−0.04	+0.05	77%	+0.001	−0.001	35
H-short	LSTM	−0.47	−0.69	+0.23	91%	−0.021	−0.031	35
	GRU	−0.39	−0.49	+0.10	71%	−0.005	−0.020	35
	RNN	−0.40	−0.47	+0.06	74%	−0.008	−0.018	35
	CNN	−0.04	−0.12	+0.08	80%	−0.002	−0.004	35
	Transformer	0.00	−0.04	+0.04	77%	0.000	0.000	35
H-long	LSTM	−0.10	−0.21	+0.11	60%	−0.031	−0.056	35
	GRU	−0.11	−0.11	0.00	46%	−0.043	+0.024	35
	RNN	−0.11	−0.13	+0.01	54%	−0.011	−0.009	35
	CNN	−0.15	−0.13	−0.02	34%	−0.006	−0.038	35
	Transformer	−0.11	−0.12	+0.02	49%	+0.108	+0.044	35
Dense	LSTM	+0.90	+0.02	+0.88	40%	−0.165	−0.048	15/35
	GRU	+0.35	−0.08	+0.43	44%	−0.025	−0.075	16/35
	RNN	−0.13	+0.02	−0.15	35%	0.000	−0.018	17/35
	CNN	−0.08	−0.20	+0.12	62%	+0.045	+0.119	13/35
	Transformer	+1.19	+0.34	+0.85	54%	−0.100	−0.004	13/35
Rare	LSTM	−0.07	−0.08	+0.01	60%	+0.326	+0.202	35
	GRU	−0.12	−0.20	+0.08	71%	+0.217	+0.103	35
	RNN	+0.01	−0.14	+0.15	80%	+0.339	+0.217	35
	CNN	+0.36	+0.11	+0.25	94%	+0.583	+0.520	35
	Transformer	+0.03	−0.12	+0.15	77%	+0.424	+0.188	35

Table 12: Full horizon ladder: median within-entity F_{rel} by forecast horizon, all architectures. h_1–h_3 are the domain-specific horizons listed in each block; h_1 = H-short and h_3 = H-long. $\Delta = F_{\text{H-long}} - F_{\text{H-short}}$.

Learner	Next	h_1	h_2	h_3	Δ(long−short)
Glucose (horizons: next, 30m, 1h, 2h)					
LSTM	0.02	0.01	0.01	0.01	0.00
GRU	0.02	0.02	0.01	0.01	−0.01
RNN	0.01	0.01	0.01	0.01	0.00
CNN	0.01	0.01	0.02	0.02	0.00
Transformer	0.04	0.05	0.04	0.05	0.00
Electricity (horizons: next, 30m, 2h, 6h)					
LSTM	0.13	0.22	0.30	0.24	+0.01
GRU	0.11	0.21	0.29	0.23	+0.03
RNN	0.09	0.18	0.24	0.12	−0.06
CNN	0.06	0.08	0.09	0.09	+0.01
Transformer	0.15	0.25	0.32	0.34	+0.09
Weather (horizons: next, 1h, 6h, 24h)					
LSTM	0.00	0.02	0.10	0.09	+0.07
GRU	−0.01	0.05	0.10	0.07	+0.02
RNN	−0.03	0.02	0.13	0.07	+0.05
CNN	0.08	0.10	0.11	0.05	−0.05
Transformer	0.03	0.04	0.08	0.22	+0.18

Table 13: Online sliding-window stress test: median relative degradation on a held-out window after online updates (negative = improvement), with the fraction of entities that degrade in parentheses (25–35 entities per domain).

Domain	Learner	Next	Dense	Rare
Glucose	LSTM	−0.88 (0%)	−0.45 (0%)	−0.62 (0%)
	GRU	−0.93 (0%)	−0.53 (0%)	−0.66 (0%)
	RNN	−0.92 (0%)	−0.49 (0%)	−0.62 (0%)
	CNN	−0.95 (0%)	−0.61 (4%)	−0.76 (0%)
	Transformer	−0.96 (0%)	−0.40 (12%)	−0.47 (12%)
Electricity	LSTM	−0.90 (0%)	−0.68 (0%)	−0.41 (31%)
	GRU	−0.91 (0%)	−0.73 (0%)	−0.47 (24%)
	RNN	−0.91 (0%)	−0.71 (0%)	−0.46 (28%)
	CNN	−0.93 (0%)	−0.78 (0%)	−0.57 (20%)
	Transformer	−0.97 (0%)	−0.73 (4%)	−0.12 (32%)
Weather	LSTM	−0.80 (0%)	−1.00 (0%)	−1.00 (0%)
	GRU	−0.90 (0%)	−1.00 (0%)	−1.00 (0%)
	RNN	−0.91 (0%)	−1.00 (0%)	−1.00 (0%)
	CNN	−0.97 (0%)	−1.00 (0%)	−1.00 (0%)
	Transformer	−0.98 (0%)	−1.00 (0%)	−0.99 (0%)

Table 14: Median within-entity relative forgetting F_{rel} for all architectures (pairwise $A \to B$ transitions). H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather). Rare events: hypo (glucose), peak (electricity), extreme (weather).

Domain	Learner	Next	H-short	H-long	Dense	Rare	n (next)
Glucose	LSTM	0.02	0.01	0.01	0.14	0.09	2791
	GRU	0.02	0.02	0.01	0.23	0.12	2839
	RNN	0.01	0.01	0.01	0.27	0.12	2846
	CNN	0.01	0.01	0.02	0.47	0.35	2946
	Transformer	0.04	0.05	0.05	0.34	0.31	2063
Electricity	LSTM	0.13	0.22	0.24	0.86	0.91	5525
	GRU	0.11	0.21	0.23	0.93	0.97	5718
	RNN	0.09	0.18	0.12	0.89	0.89	5752
	CNN	0.06	0.08	0.09	0.80	1.20	5745
	Transformer	0.15	0.25	0.34	1.03	1.31	5425
Weather	LSTM	0.00	0.02	0.09	1.92	1.41	3309
	GRU	−0.01	0.05	0.07	2.15	1.70	5211
	RNN	−0.03	0.02	0.07	1.98	1.45	5491
	CNN	0.08	0.10	0.05	2.49	2.05	6630
	Transformer	0.03	0.04	0.22	2.19	1.80	2398

Table 15: Median savings ratio $1 - T_{\text{relearn}}/T_{\text{scratch}}$ after within-entity chains, all architectures ($n = 35$ entities per domain). Higher values mean faster relearning; negative values mean relearning took longer than training from scratch. H-short/H-long: 30m/2h (glucose), 30m/6h (electricity), 1h/24h (weather).

Domain	Learner	Next	H-short	H-long	Dense	Rare
Glucose	LSTM	0.95	0.94	0.67	0.94	0.95
	GRU	0.94	0.93	0.67	0.94	0.95
	RNN	0.92	0.93	0.50	0.96	0.96
	CNN	0.88	0.80	0.00	0.78	0.81
	Transformer	0.80	0.50	−1.62	0.00	−1.00
Electricity	LSTM	0.96	0.94	0.75	0.91	0.73
	GRU	0.94	0.92	0.83	0.89	0.55
	RNN	0.93	0.94	0.75	0.86	0.68
	CNN	0.57	0.67	0.50	−0.33	−0.40
	Transformer	0.57	0.50	−0.45	−0.23	−2.44
Weather	LSTM	0.97	0.97	0.91	0.92	0.58
	GRU	0.97	0.98	0.75	0.85	0.50
	RNN	0.97	0.97	0.82	0.80	−0.21
	CNN	0.79	0.89	0.70	0.87	0.31
	Transformer	0.83	0.86	0.06	−0.02	−4.33

Table 16: Event-threshold sensitivity (LSTM, 20 entities, 80 pairwise transitions). Weather seasonal tail events use paired upper/lower quantiles. At $q = 0.99/0.01$ weather seasonal tail events are so rare that the loss saturates and F_{rel} collapses toward -1; this value is an artifact, not recovery.

Task	Threshold	Entities	Pairs	F_{rel}
Glucose (hypo)	60 mg/dL	20	80	+0.03
	70 mg/dL	20	80	+0.01
	80 mg/dL	20	80	−0.01
Electricity (peak)	$q = 0.90$	20	80	+0.04
	$q = 0.95$	20	80	0.00
	$q = 0.99$	20	80	−0.16
Electricity (next, control)	–	20	80	−0.07
Weather (extreme)	$q = 0.90$ / 0.10	20	80	+0.20
	$q = 0.95$ / 0.05	20	80	−0.02
	$q = 0.99$ / 0.01	20	80	−1.00

Table 17: Event forgetting tracks label-prior (base-rate) shift. For every valid within-entity pair $(A \to B)$ we compute the relative forgetting of a predictor that knows only the base rate, $F_{\text{rel}}^{\text{prior}} = [\text{BCE}(y_A; \bar{p}_B) - \text{BCE}(y_A; \bar{p}_A)]/\text{BCE}(y_A; \bar{p}_A)$, with $\bar{p}$ the training-split positive rate and y_A the test labels of A. n: valid pairs (Pooled: all learners). ρ: Spearman correlation between $F_{\text{rel}}^{\text{prior}}$ and observed F_{rel}; $\rho_{|W}$: partial Spearman controlling for the Wasserstein distance between A and B inputs; ρ_{abs}: the same correlation on unnormalized BCE differences (pooled row: median over learners). Rare events: hypo (glucose), peak (electricity), extreme (weather). Labels were rebuilt from the pairwise pipeline and reproduce each pair's stored constant-class baseline exactly. All correlations have $p < 10^{-10}$.

		Rare event				Dense event			
Domain	Learner	n	ρ	$\rho_{\|W}$	ρ_{abs}	n	ρ	$\rho_{\|W}$	ρ_{abs}
Glucose	LSTM	2,753	0.50	0.49	0.45	3,021	0.37	0.22	0.36
	GRU	2,762	0.47	0.46	0.43	3,035	0.36	0.22	0.36
	RNN	2,795	0.46	0.45	0.41	3,041	0.36	0.21	0.35
	CNN	2,876	0.43	0.42	0.39	3,046	0.33	0.17	0.33
	Transformer	2,877	0.44	0.43	0.42	3,050	0.37	0.21	0.37
	Pooled	14,063	0.46	0.44	0.42	15,193	0.35	0.20	0.36
Electricity	LSTM	4,376	0.86	0.85	0.82	5,960	0.63	0.59	0.64
	GRU	4,393	0.86	0.84	0.82	5,964	0.62	0.58	0.64
	RNN	4,362	0.85	0.83	0.81	5,961	0.63	0.59	0.64
	CNN	4,635	0.86	0.84	0.80	6,003	0.60	0.55	0.60
	Transformer	4,519	0.84	0.82	0.80	6,005	0.62	0.58	0.63
	Pooled	22,285	0.85	0.84	0.81	29,893	0.62	0.58	0.64
Weather	LSTM	4,666	0.72	0.73	0.65	4,716	0.87	0.77	0.82
	GRU	4,770	0.70	0.71	0.62	4,786	0.86	0.76	0.81
	RNN	4,866	0.71	0.73	0.63	4,851	0.87	0.78	0.83
	CNN	5,081	0.71	0.73	0.61	4,965	0.88	0.76	0.81
	Transformer	5,035	0.70	0.71	0.65	5,110	0.87	0.78	0.81
	Pooled	24,418	0.71	0.72	0.63	24,428	0.87	0.77	0.81

Table 18: Per-segment positive rates of the event objectives on the monthly chain segments ($n = 35$ entities per domain; same thresholds, horizons and segmentation as training). Pooled: overall positive rate; P10–P90: spread of per-segment rates; $< 2\%$: fraction of segments that are almost event-free; Seg. 0: median rate of the first chain segment (January for electricity and weather); Range: median within-entity max$-$min monthly rate. Last two columns: median base-rate-only vs. observed relative forgetting over all within-entity pairs and learners (Table 17). Glucose hypoglycaemia is the only rare event with a seasonally stable base rate; electricity peaks concentrate in summer, and weather extremes (cold and heat tails of a year-round quantile) are common in winter and summer but nearly absent in spring and autumn.

Domain	Event	Pooled	P10–P90	$< 2\%$	Seg. 0	Range	$F_{\text{rel}}^{\text{prior}}$	F_{rel}
Glucose	Rare (hypo)	0.07	0.01–0.15	19%	0.07	0.09	0.01	0.19
	Dense	0.53	0.41–0.64	0%	0.50	0.35	0.01	0.28
Electricity	Rare (peak)	0.06	0.00–0.20	53%	0.09	0.34	0.22	1.05
	Dense	0.55	0.37–0.68	0%	0.63	0.48	0.02	0.90
Weather	Rare (extreme)	0.16	0.00–0.40	23%	0.40	0.64	0.06	1.66
	Dense	0.57	0.04–1.00	7%	0.01	1.00	0.42	2.13

Table 19: Reference swap (P5; LSTM, within-entity transitions sampled from the Table 1 cohort, same seed per transition under both conditions). Dense: event threshold at the entity median (paper) vs. the median of each segment's own training split (label base rate approximately fixed across segments). Rare: entity $q_{0.95}$ (peak) or $q_{0.05}/q_{0.95}$ (extreme) vs. the same quantiles of each segment's own training split; glucose hypoglycemia uses a clinical threshold and is not swapped. Next: per-segment (paper) vs. per-entity z-normalization (between-segment level shifts reach the target). F_{rel} columns: medians; Δ: median paired difference (swapped − paper); Swap >: fraction of transitions with higher forgetting after the swap; p: Wilcoxon signed-rank. Skill: median pre-transfer skill_A. Skill is matched in every row except weather rare, where segment-relative extremes are harder to predict; on the 43 transitions with skill matched within 0.1, F_{rel} still falls from 1.43 to 0.06.

		F_{rel}					Skill_A	
Domain	n	Paper	Swapped	Δ	Swap >	p	Paper	Swapped
Dense: entity → segment median								
Glucose	398	0.13	−0.05	−0.17	25%	$< 10^{-10}$	0.64	0.64
Electricity	398	0.93	0.16	−0.59	22%	$< 10^{-10}$	0.74	0.71
Weather	358	2.31	0.06	−1.96	19%	$< 10^{-10}$	0.48	0.50
Rare: entity → segment quantiles								
Electricity	392	1.13	0.13	−0.80	28%	$< 10^{-10}$	0.53	0.53
Weather	277	1.35	0.12	−1.04	32%	$< 10^{-10}$	0.47	0.19
Next: segment → entity norm.								
Glucose	392	0.01	0.02	0.00	53%	0.12	0.42	0.42
Electricity	389	0.18	0.26	+0.05	62%	$< 10^{-10}$	0.18	0.18
Weather	309	0.07	0.59	+0.43	73%	$< 10^{-10}$	0.47	0.44

Table 20: Recalibration of the weather-extreme recurrence gap (35 cities per learner; recurrence month r = January, 12 months after $\mathcal{S}_1$). The chained model is recalibrated on r's training split and scored on its test split. Median positive rate: 0.43 in r vs. 0.19 in the preceding month; $\bar{\hat{p}}_{\mathrm{chain}}$: median mean predicted probability of the chained model on r. Closure columns report $1 - \operatorname{med} G^{\mathrm{cal}}_{\mathrm{rec}} / \operatorname{med} G_{\mathrm{rec}}$ for a logit offset (Bias), the same offset fitted on the first 168 training windows (≈1 week), and an affine logit map (Platt); values above 100% mean the recalibrated chain beats the scratch model. Last column: median gap remaining after Platt recalibration.

				Gap closed			
Learner	n	$\bar{\hat{p}}_{\mathrm{chain}}$	G_{rec}	Bias	Bias (1 wk)	Platt	$G^{\mathrm{cal}}_{\mathrm{rec}}$ (Platt)
LSTM	35	0.10	0.25	115%	101%	117%	−0.04
GRU	35	0.14	0.25	92%	81%	110%	−0.02
RNN	35	0.12	0.27	63%	62%	66%	+0.09
CNN	35	0.10	0.76	64%	60%	86%	+0.11
Transformer	35	0.15	0.32	44%	50%	66%	+0.11
Pooled	175	0.12	0.33	72%	69%	77%	+0.07

Table 21: Where the damage lives: weather extremes, first-segment test set, before vs. after the within-entity chain (medians over $n = 35$ cities; PM: parameter-matched transformer). Before/After: BCE of the full model. CKA: linear CKA of first-segment features. Probe AUROC: linear probe trained on frozen features before and after the chain. Transplant: fraction of the lost loss recovered by restoring the pre-chain output head on post-chain features; Retrain: by re-initializing the head and retraining it alone on the first segment. Freeze/Replay: post-chain BCE when the encoder is frozen during the chain, or with tiny replay of first-segment windows. For the LSTM the representation still separates events and the loss is recovered by the head alone, while freezing does not help; for the transformer, probe accuracy drops, the original head no longer fits, and protecting the representation (freeze, replay) reduces forgetting.

	BCE				Head recovery		BCE with	
Learner	Before	After	CKA	Probe AUROC	Transplant	Retrain	Freeze	Replay
LSTM	0.47	0.68	0.68	0.92→0.94	61%	102%	0.81	0.66
Transformer	0.43	0.83	0.55	0.91→0.85	−2%	85%	0.63	0.63
Transformer (PM)	0.39	0.65	0.68	0.92→0.89	9%	55%	0.55	0.66