ARCHITECTURAL LEAKAGE: PER-TIMESTEP SEQUENCE MODELS THAT QUIETLY SEE THE FUTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Leakage in time-series evaluation is usually a data problem: a split that lets training see the test period, or a feature computed after the label. We document a form those checks cannot see, *architectural leakage*, in which the split and the features are clean and the model itself has a path from future inputs to past outputs. It arises whenever a per-timestep model conditions on a whole-sequence quantity—a bidirectional encoder, unmasked attention, symmetric padding, a pooled context vector—and it is invisible to split- and feature-level checks because the model is trained and scored under the same protocol.

We give a model-agnostic detector, the future-perturbation score (FPS): perturb timesteps after k, measure the change at timesteps $\leq k$; a causal model scores exactly zero. It needs only `forward(x, t)`: a replay on the truncated prefix detects the leak, and four value perturbations of the future narrow down what is read; a shuffle-only test would pass the form in the audited model, a mean-pooled context, because a mean is permutation-invariant. We then measure what such a leak is worth: on a monotone early-warning label, an oracle b-bit channel carrying the onset admits a label-only score computed from the labels alone; on PhysioNet2019 a *single* such bit reaches 0.957 (at least 0.979 under a validation-selected partition), above every causal model we can train, and a bit predicted from whole-record summaries of the inputs already reaches 0.86. An audit of matched causal/non-causal pairs from four families (LSTM/BiLSTM, causal/symmetric TCN, masked/unmasked Transformer, running/whole-sequence pooling) finds every causal member of the four at 0.76–0.80 and every non-causal member beating its causal twin on all twenty splits, up to 0.974; the gap survives per-patient scoring and the challenge utility. A routed additive model reported at $+14.2$ AUROC over its causal baseline serves as the case study: a fair comparison gives -0.3, and reading its measurement process causally recovers $+1.7$ AUROC on 19 of 20 splits ($+0.02$ normalised utility)—most of it without any router. The detector (under a minute per checkpoint) and the label-only scores (minutes on a CPU) are released.

1 INTRODUCTION

A model that predicts sepsis at hour k must not read hour $k{+}1$. Every practitioner knows this, and the standard defences—chronological splits, features computed only from the past, no target-derived inputs (Kaufman et al., 2012; Kapoor & Narayanan, 2023)—are aimed at the data. This paper is about a leak the data-side defences cannot see, because it is not in the data. It is in the model.

Any per-timestep sequence model that conditions its output at step k on a quantity computed from the whole sequence has a path from $x_{>k}$ to $\hat{y}_k$. Bidirectional recurrence, attention without a subsequent mask, convolution with symmetric padding, and a context vector pooled over the sequence all have one. When an output is needed only after the whole record exists—a many-to-one label, or offline annotation of a finished stay—this is harmless. When each output must be issued at its own timestep, as in early warning, it is not, and nothing in the training or evaluation protocol will object: the loss is computed at every step, the metric is computed at every step, and the model is scored under exactly the conditions it was fit under. The number goes up, the paper is written, and the leak ships.

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.

We call this *architectural leakage*. It breaks the same rule as temporal leakage—no output may depend on information unavailable when it is issued—and the two-run test for that rule is not ours (Fonseca, 2026; Saggese & Smith, 2025). What is new is where the violation sits (inside a trained model whose split and features are clean), a benchmark of which cheaper checks miss which leaks (a shuffle-only check passes a mean-pooled context), and a measurement of what the violation is worth. That a bidirectional encoder cannot be run online is textbook; what has not been established is that on a monotone label the misuse is worth more than the entire measured record, and that a shuffle-only version of the test passes a mean-pooled context. Our entry point was one specific model—a routed additive time-series model whose reported $+14.2$ AUROC over its causal baseline turned out to be entirely this artefact—but the phenomenon is not specific to it.

Contributions.

1. **A detector.** The future-perturbation score (§4): perturb timesteps after k, measure the change at timesteps $\leq k$. A causal model scores exactly zero; the test needs only `forward(x, t)` (inputs and each sequence's last valid index) and under a minute per checkpoint. A prefix replay detects; shift, shuffle, scale and warp diagnose what is read; shuffle alone is blind to a mean-pooled context.
2. **A label-only reference.** On a monotone early-warning label, a b-bit sequence-level side channel admits a score that conditions on the onset and nothing else (§5). Its AUROC follows from the label distribution alone. On PhysioNet2019 one oracle bit reaches 0.957, more than every measured variable read causally. A trained model's channel carries only what the record reveals, so we also measure the input-recoverable score: a bit predicted from whole-record summaries reaches 0.86, four bits 0.90–0.92 (0.917 from stay length alone). The oracle score measures how exposed the label is, not how a given model will do.
3. **A matched-pair audit.** Causal/non-causal pairs from four families, differing in one line of code each (§6). The causal column is flat (0.76–0.80); every non-causal member gains on all twenty splits, from $+6.2$ (symmetric TCN, whose receptive field bounds its reach) to $+19.2$ (BiLSTM). Every non-causal model fails the detector; every causal one passes.
4. **A checklist** (§7), a released test, and our failed hypotheses in Appendix K rather than omitted.
5. **A case study, and the part of it that was real.** The routed additive model is dissected in §8 and Appendix C: what the channel carries (about half is recoverable from the measurement process alone, a quarter of that causally), why capping the gate fails, and what a corrected benchmark shows. Reading the measurement process causally recovers $+1.7$ AUROC on 19 of 20 splits, most of it without any router (§8.1), while a fair version of the original comparison gives -0.3.

2 Related work

Leakage. Kaufman et al. (2012) formalised leakage as the use of information at training time that is unavailable at prediction time, and catalogued its data-side forms: target-derived features, and train/test contamination. Kapoor & Narayanan (2023) surveyed seventeen fields and found it the leading cause of irreproducible machine-learning claims, again at the level of splits and features, as does the REFORMS standard that followed (Kapoor et al., 2024). None has a category for the model itself. Fonseca (2026) formalises look-ahead-freedom as temporal non-interference and checks it dynamically, perturbing only future data and recomputing the decision; Saggese & Smith (2025) expose future-peeking as a mismatch between batch and streaming execution. FPS is that two-run test applied to trained neural sequence models, with a replay on the cropped prefix as the detector; what we add are value perturbations that diagnose what a leaking model reads, a benchmark of which checks miss which leaks (Table 3), and its use as an audit instrument. Here the split respects sequence boundaries and no feature is target-derived, yet the model reads the future, because its architecture computes a function of the whole sequence and feeds it to every step.

Sequence families and their causal variants. Each family we audit has a causal and a non-causal member that are a documented one-line switch apart: LSTM and BiLSTM (Hochreiter & Schmidhuber, 1997; Graves & Schmidhuber, 2005); temporal convolution with left versus symmetric

padding (Bai et al., 2018); attention with and without the subsequent mask (Vaswani et al., 2017). That they are inappropriate for per-timestep prediction is understood in principle; what has not been measured is how much they gain when misapplied, and how that gain compares with what the label alone allows.

The case-study family. Generalised additive models decompose the prediction into per-feature terms, so the attribution is the model rather than an approximation of it; Neural Additive and Neural Basis Models (Agarwal et al., 2021; Radenovic et al., 2022) are the neural revival; instance-dependent shape functions are classical (Hastie & Tibshirani, 1993) and recently revisited for tabular data (Richman & Wüthrich, 2023; Xiong et al., 2026). The model we audit (Anonymous, 2026) carries the latter's design—several experts per feature and a dynamic gate—to sequences. **We do not claim the architecture as a contribution.** The relevant difference is the setting: a tabular row has no time axis, so pooling raises no causality question; a statistic pooled over a *sequence* and evaluated *per timestep* silently stops being causal. Chen et al. (2022) show a router should help when the problem has cluster structure a single expert cannot capture; our case study is the empirical complement on a domain where that condition appears weak.

Leakage in clinical prediction. Wong et al. (2021) externally validated a deployed sepsis model across $38{,}455$ hospitalisations and found AUC 0.63, far below its published performance. A recurring cause is that models learn the care process rather than the physiology: that a laboratory value was ordered is itself a signal of clinician concern (Agniel et al., 2018). Architectures that model missingness explicitly (Che et al., 2018) treat this as information to be used; Appendix C.2 shows the opposite hazard, in which the measurement process enters through a gate where no additive attribution can see it. Protocol choices are known to move ICU results substantially (van de Water et al., 2024; Fleuren et al., 2020; Moor et al., 2021), and a model that reads the care process instead of physiology is shortcut learning (Geirhos et al., 2020).

3 SETUP

Per-timestep prediction and causality. A sequence $x_{1:T}$ with $x_k \in \mathbb{R}^F$ carries a label y_k at every step, issued at step k. A model is *causal* if $\hat{y}_k$ is a function of $x_{1:k}$ alone—the information available when $\hat{y}_k$ is issued. Every early-warning deployment is causal by construction—the future has not happened—so a per-timestep benchmark of an early-warning task measures a model only if the model is causal too; offline annotation of a finished record may read it whole, and we make no claim about it. The primary benchmark is PhysioNet2019 (Reyna et al., 2020) (sepsis onset, $F{=}40$, per-timestep AUROC), whose label is monotone: once positive, a patient stays positive.

The case-study backbone. We study the GATSM family (Kim et al., 2026): a pointwise per-feature basis expansion, a linear projection, sinusoidal position encoding, and additive-scoring multi-head attention with causal masking, followed by a linear read-out. Every stage is causal: the basis is applied per timestep and the attention masks the strict upper triangle. Feature-wise independence is enforced by grouped convolutions, which is what makes the additive decomposition exact.

Routed variant. The routed model under audit (Anonymous, 2026)—a manuscript under review, not a published result, so we call its numbers *reported*—replaces the single combiner with C expert combiners per feature and a router. The router reads

$$\text{context} \;=\; \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \phi(x)_t, \qquad \mathcal{T} = \{t : t \le T_{\text{valid}}\}, \tag{1}$$

where T_{valid} is the last unpadded timestep, produces per-feature logits over experts, applies a top-m mask and a softmax, and mixes the expert outputs. Equation (1) is the entire subject of this paper.

Tasks and protocol. Ten benchmarks: eight UCR/UEA/Monash series, PhysioNet2012 (Silva et al., 2012) (in-hospital mortality, many-to-one) and PhysioNet2019 (Reyna et al., 2020) (sepsis onset, many-to-many, scored as per-timestep AUROC). Only the last is many-to-many, which is why Appendix G adds a controlled pair. Splits are 60/20/20, re-drawn per seed, five seeds (0–4) unless stated, preprocessing fit on train only. Throughout, $\pm$ is the population standard deviation (ddof$=0$)

over the seeds. We report paired comparisons; with five seeds a Wilcoxon signed-rank test bottoms out at $p = 0.062$.

4 ARCHITECTURAL LEAKAGE, AND A DETECTOR

4.1 THE FUTURE-PERTURBATION SCORE

Let f map a batch of sequences to per-timestep outputs and fix a cut k. For a perturbation π that alters only timesteps $> k$, define

$$\mathrm{FPS}_\pi(f) \;=\; \max_{j \le k} \frac{\mathbb{E}\,\big|f(x)_j - f(\pi(x))_j\big|}{\mathrm{sd}\, f(x)}, \qquad \mathrm{FPS}(f) = \max_\pi \mathrm{FPS}_\pi(f), \tag{2}$$

over five perturbations: *shift* adds a constant to every future timestep, *shuffle* permutes them, *scale* multiplies them by a constant, *warp* applies a fixed strictly increasing non-linear map ($x \mapsto \mathrm{sign}(x)|x|^{1.5}$), and *trunc* crops the record after k and sets `t` to k—the two-run prefix replay. The normaliser is the output's standard deviation, so the score ignores an additive offset or rescaling of the output. Scored in float64, every causal checkpoint we audited gives exactly 0 (Appendix A); we flag $\mathrm{FPS} > 10^{-6}$, and below it report 'no leak detected', not 'causal'. Detection rests on the prefix replay, the two-run definition itself, so it catches all fourteen synthetic leaks of Table 3 by construction (two forward passes, no false positive). The point is which cheaper checks miss: shuffle alone catches four and passes a mean-pooled context; zero-padded truncation misses a model reading the tensor's extent. The value perturbations diagnose: which fire and which stay silent constrains what is read (a silent shuffle with a firing shift means an order-invariant statistic such as a mean). Zero is evidence, not a certificate. The test uses only `forward(x, t)` (`t` is each sequence's last valid index), a few held-out batches and one forward pass per perturbation, under a minute per checkpoint (Appendix A). A non-zero Jacobian $\partial \hat{y}_k / \partial x_j$, $j > k$, also proves a path, but a zero one at sampled inputs does not prove its absence; FPS is the black-box counterpart that also covers hard and straight-through gates and inference-only checkpoints.

4.2 THE CASE-STUDY MODEL FAILS IT

Measured on trained PhysioNet2019 checkpoints (seed 0, $k = 24$; mean absolute change in the output logit over held-out sequences, maximised over $j \le k$), the additive baseline (GATSM) and the routed model with a running-mean basis context (the causal control of Appendix C.1) both change by exactly 0.000; the routed model as reported changes by 1.353. Figure 1 plots the full response: the two causal models are flat at zero until the perturbation boundary and rise only after it; the reported router is displaced by more than one logit across the *entire* pre-boundary region, because its gate is computed from a mean over the whole sequence. Zero is the only value consistent with causality; and the test also works on an untrained model (Appendix A). In the units of Equation (2) the reported model scores $\mathrm{FPS}_{\text{trunc}} = 1.3$, $\mathrm{FPS}_{\text{shift}} = 0.84$, $\mathrm{FPS}_{\text{scale}} = 0.25$, $\mathrm{FPS}_{\text{warp}} = 0.15$ and $\mathrm{FPS}_{\text{shuffle}} = 0$ (means over twenty checkpoints, Table 1); a shuffle-only check would have passed it.

5 ONE LEAKED BIT BEATS EVERY CAUSAL MODEL WE TRAINED

Section 4 establishes *that* a pooled context leaks. This section asks how much a leak of a given size can be worth. We answer twice: for an oracle channel that carries the onset, from the labels alone, and for a channel that must infer the onset from the record, from the inputs.

Setting. An early-warning label is *monotone*: for each sequence there is an onset $\tau \in \{1, \dots, T\} \cup \{\infty\}$ with $y_k = \mathbb{1}[k \ge \tau]$, and $\tau = \infty$ for sequences that never turn positive. A per-timestep score s_k is *causal* if s_k is a function of $x_{1:k}$ alone. A non-causal component is any function $g(x_{1:T})$ of the whole record that the score may also read; we call it a *sequence-level side channel* and measure its capacity in bits, $b = \log_2 |\mathrm{range}(g)|$. A pooled routing context has b unbounded; a shared one-hot gate over C experts has $b = \log_2 C$. Capacity bounds what g can carry, not what the record lets it carry.

The danger of such a channel on a monotone task is not that it carries physiology; it is that the cheapest thing to carry through it is τ itself.

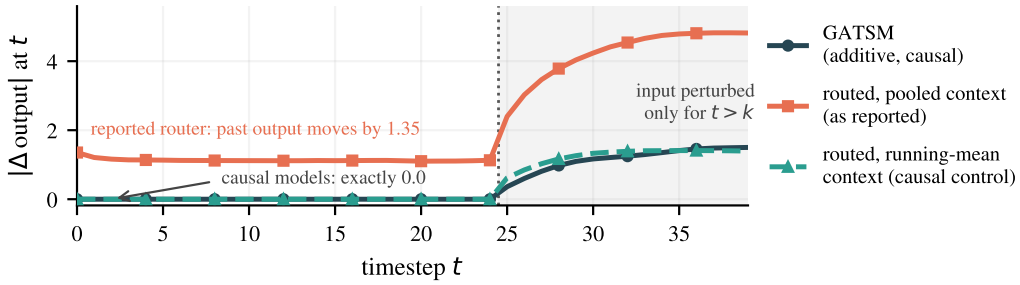

Figure 1: Measured future-perturbation response on trained PhysioNet2019 checkpoints. Only timesteps $t > k$ are perturbed (shaded). A causal model must be zero to the left of the boundary; the routed model as reported is displaced by more than one logit throughout that region.

Remark 1 (Label-only score of a b-bit channel). Fix a distribution over (T, τ) and let $\mathcal{C}_b$ be any partition of $\{1, \dots, T\} \cup \{\infty\}$ into at most 2^b cells. A b-bit channel that transmits the cell $c(\tau) \in \mathcal{C}_b$ and nothing else admits the score $s_k = \Pr[y_k = 1 \mid c(\tau), k]$, which maximises AUROC among all scores that are functions of $(c(\tau), k)$. Its AUROC, A_b, is fixed by the law of (T, τ) and the partition alone: no input enters. For $b = 0$ (no channel) it reduces to $\Pr[y_k = 1 \mid k]$, the best score that uses no measurement at all; for $b \to \infty$ it reaches 1 because $c(\tau) = \tau$ determines every y_k exactly.

The remark is elementary[1]—it is the statement that a score which conditions on the onset needs nothing else—but its value on a real label distribution is not obvious, and that is what matters. We computed A_b on PhysioNet2019 with the paper's own 60/20/20 splits, one cell reserved for $\tau = \infty$ and the rest quantile bins over the septic onsets of the training fold (Figure 2). This is one partition, so the computed A_b is *achievable*: some b-bit channel attains it using no input at all. It is therefore a lower bound on $A_b^\star = \sup_{\mathcal{C}_b} A_b$, the optimum over partitions, and an upper bound on nothing; every claim below needs only that a b-bit leak *can* reach A_b. For $b = 1$ we searched the threshold partitions: choosing the split on a validation fold gives 0.979 ± 0.002, separating early onsets from late ones and from never—2.2 points above the partition we report, so $A_1^\star \geq 0.979$, further still above every causal model. For $b = 2$ the same search over two thresholds gains only 0.2 points (0.989 ± 0.001 against the quantile partition's 0.987): past one bit, the partition we report is already close to the best threshold partition.

What a real channel recovers. The remark hands the channel the oracle cell $c(\tau)$; a trained model's channel $g(x_{1:T})$ carries only what the record reveals about τ, and nothing if $x \perp \tau$. We therefore also computed $\hat{A}_b$: the same score with $c(\tau)$ replaced by a cell predicted from whole-record summaries of the inputs (variable means, recording rates, stay length; a boosted probe cross-fitted on the training fold; Appendix B). The record reveals much of the bit: $\hat{A}_1 = 0.858 \pm 0.004$, $\hat{A}_4 = 0.901 \pm 0.003$ (dashed in Figure 2). By source, past the first bit the record's end dominates: stay length alone gives $\hat{A}_4 = 0.917$ (logistic; 0.889 boosted), clinical values 0.794, recording rates 0.835, both 0.852 ($b = 1$: length and clinical with rates tie at 0.81); the ICU offset known at the record's start gives no more than A_0. Trained channels carry as much: the future-derived part of a model's first output alone—one number per record, a lower bound—gives $\hat{A}_4 = 0.89$ for the routed model, 0.90 for the pooled LSTM, 0.83/0.82 for BiLSTM/unmasked attention and 0.68 for the 3-step TCN (Table 4). A_b measures how exposed the label is; neither score bounds a trained model, which also reads the causal measurements.

Three readings. *(i) The benchmark is one bit from solved.* With no measurement, the time index alone scores 0.703. Forty clinical variables, read causally by the best of the architectures we trained,

[1] $s_k = \Pr[y_k{=}1 \mid c(\tau), k]$ is the posterior given the information $(c(\tau), k)$, and the posterior—any strictly increasing transform of the likelihood ratio—maximises AUROC (ties scored $1/2$) among all scores that are functions of the same information (Neyman–Pearson). Its AUROC is a functional of the joint law of $(y_k, c(\tau), k)$, which is fixed by the law of (T, τ) and the partition. At $b = 0$ the cell is trivial; as $b \to \infty$ the partition separates every onset, and $y_k = \mathbb{1}[k \geq \tau]$ is then determined.

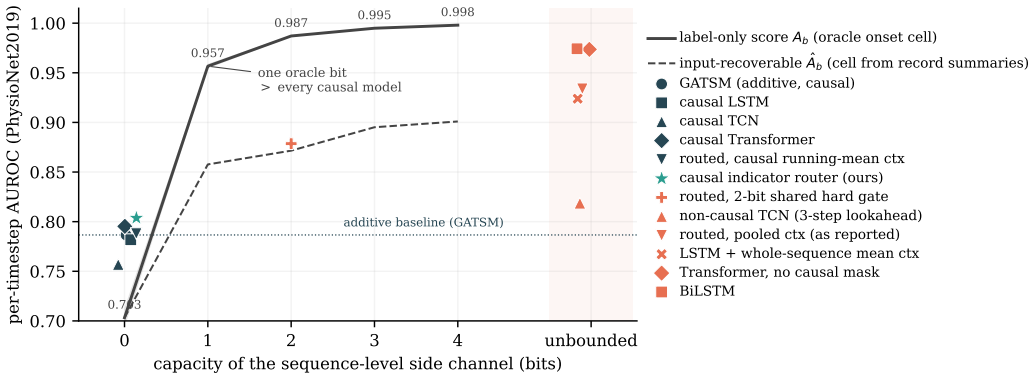

Figure 2: The label-only score A_b on PhysioNet2019 (5 split seeds for the curve, twenty for every trained marker except the causal running-mean routed model; band is one standard deviation). The curve uses no input whatsoever—only the onset time binned into 2^b cells. Every trained model from the audit is placed at the nominal channel capacity its architecture permits; capacity is not what the channel recovers (§5, 'What a real channel recovers'). Causal models span 0.757–0.804; a *single* leaked bit reaches 0.957 (≥ 0.979 with a validation-selected split). Models with an unrestricted future path reach 0.974, those with a single pooled vector 0.924–0.934.

add at most $+0.10$ (to 0.76–0.80). One bit of "will this patient turn septic, and roughly when" adds $+0.25$. On this monotone label, the causal models we trained extract less from the whole record than an oracle's first bit carries.

(ii) Capping the gate is not enough. A two-bit oracle channel admits 0.987 and two bits inferred from record summaries 0.872–0.877; the two-bit shared hard gate of Section 8.2 reached 0.879, close to $\hat{A}_2$—a descriptive match only, since that model also reads the causal inputs.

(iii) Family matters less than access. The right-hand column of Figure 2 holds an additive, a recurrent and an attention model. Their causal versions differ by at most 1.3 points (0.782–0.795); their non-causal versions (0.924–0.974) are every one above every causal number. What separates the columns is whether $g(x_{1:T})$ exists. Capacity in bits does not rank the unbounded channels within the non-causal column, and we do not claim it does.

6 A MATCHED-PAIR AUDIT OF EARLY-WARNING ARCHITECTURES

If Section 5 is right, the gain from a future-reading channel is not a property of routing, of additivity, or of any particular design. It is a property of any per-timestep model that has a path from $x_{>k}$ to $\hat{y}_k$. We test this by training *matched pairs* from four families that differ in exactly one line of code (Table 1).

family	causal member	non-causal member	the one line
recurrent	LSTM	BiLSTM	`bidirectional=True`
convolutional	TCN, left padding	TCN, symmetric padding	`pad=(kernel-1)/2`
attention	Transformer, causal mask	Transformer, no mask	`mask=None`
pooled context	LSTM $\oplus$ running mean	LSTM $\oplus$ sequence mean	`cumsum` $\rightarrow$ `mean`

All eight models share nominal width (64; `bidirectional` doubles the BiLSTM's output width, Appendix L), depth (2), dropout (0.1), optimiser (AdamW, 10^{-3}), batch size (64), early stopping (patience 20 on validation loss) and twenty $60/20/20$ splits. The fourth family is the routed model's context path with the experts removed, so that the mechanism of Section 4 can be compared with the others on equal terms. Each checkpoint is scored twice: per-timestep AUROC on the held-out fold, and the future-perturbation score (FPS) of Section 4, which is zero for a causal model.

Table 1: Matched causal / non-causal pairs on PhysioNet2019 (per-timestep AUROC). Every row is twenty splits. FPS is the future-perturbation score of Equation (2) under each perturbation, averaged over the twenty non-causal checkpoints of each row; every causal checkpoint we scored (125: 80 causal pair members, 20 GATSMs, 20 causal indicator routers, 5 causal running-mean routers) is exactly 0 under every perturbation (scored in float64) and is omitted. The gap is the paired difference in AUROC points and the number of splits on which the non-causal member wins; every $20/20$ has exact Wilcoxon $p < 10^{-4}$.

causal member		non-causal member							gap	
model	AUROC	model	AUROC	$\text{FPS}_{\text{shift}}$	FPS_{shuf}	$\text{FPS}_{\text{scale}}$	FPS_{warp}	$\text{FPS}_{\text{trunc}}$	pts	wins
LSTM	0.782 ± 0.011	BiLSTM	0.974 ± 0.001	7.7e−1	3.3e−2	3.1e−1	9.0e−2	2.3e0	+19.2	20/20
TCN, left pad	0.757 ± 0.012	TCN, sym. pad	0.819 ± 0.011	2.5e0	2.6e−1	3.1e0	3.9e−1	2.9e0	+6.2	20/20
Transf., masked	0.795 ± 0.010	Transf., unmasked	0.974 ± 0.002	4.8e−1	6.8e−2	2.6e−1	2.2e−1	2.2e0	+17.8	20/20
LSTM+run. mean	0.781 ± 0.009	LSTM+seq. mean	0.924 ± 0.005	1.7e0	0	7.0e−1	1.5e−1	1.2e0	+14.2	20/20
GATSM (additive)	0.787 ± 0.010	routed, pooled	0.934 ± 0.006	8.4e−1	0	2.5e−1	1.5e−1	1.3e0	+14.7	20/20

Table 2: Per-patient AUROC ($\max_t p_t$) and Challenge-2019 normalised utility (validation threshold) for the pairs of Table 1 and the indicator router; ddof$= 0$.

		per-patient AUROC			normalised utility		
pair (causal / non-causal)	n	causal	non-causal	gap	causal	non-causal	gap
LSTM / BiLSTM	20	0.778 ± 0.017	0.907 ± 0.005	+12.8	0.316 ± 0.021	0.693 ± 0.013	+0.38
TCN, left / sym. pad	20	0.769 ± 0.014	0.892 ± 0.011	+12.3	0.291 ± 0.020	0.409 ± 0.018	+0.12
Transf., masked / unmasked	20	0.809 ± 0.011	0.923 ± 0.006	+11.3	0.337 ± 0.017	0.732 ± 0.012	+0.39
LSTM + run. / seq. mean	20	0.778 ± 0.016	0.866 ± 0.009	+8.8	0.315 ± 0.021	0.605 ± 0.020	+0.29
GATSM / routed, as reported	20	0.775 ± 0.018	0.885 ± 0.011	+10.9	0.330 ± 0.023	0.634 ± 0.019	+0.30
GATSM / causal indicator router (ours)	20	0.775 ± 0.018	0.796 ± 0.013	+2.0	0.330 ± 0.023	0.353 ± 0.017	+0.02

The causal column is flat. Four families, four causal models, all within 0.757–0.795 over twenty splits. The additive baseline is not handicapped by its interpretability constraint: it trails the best causal model in the table, the masked transformer, by 0.8 points, and that transformer leads the LSTM by 1.3—small beside the gaps in the right-hand column. The causal running-mean context adds nothing to the LSTM (0.78 in both rows), consistent with the running-mean basis router of Appendix C.1 (-0.27).

The non-causal column tracks access, not family. BiLSTM and unmasked attention—an unrestricted path from every future timestep to every past output—both land at 0.974. The pooled-context model, whose channel is one mean vector, lands at 0.924, and the routed model as reported, the same vector read by a router, at 0.934. The symmetric-padding TCN, whose look-ahead is bounded at 3 steps, gains the least ($+6.2$). Bounded look-ahead (Table 10) gains about 3 points within six hours and 11 (pooled) or 15 (Transformer) at 24: most of the gain lies beyond the label's six-hour horizon, a range holding both onset and, for many stays, the record's end.

The FPS column is diagnostic. Every causal model scores exactly 0 (float64); every non-causal one scores at least 0.4, more than five orders of magnitude above the 10^{-6} flag. Shuffle alone underreports the BiLSTM by a factor of 23 against shift and misses the pooled-context model entirely (exactly 0 against 1.7 under shift): a mean is permutation-invariant, so the family that most resembles the case-study model is the one a shuffle-only check clears.

The gap is not a pooling artefact. Pooled per-timestep AUROC weights long stays heavily and could, in principle, exaggerate a leak. Table 2 scores the same checkpoints once per stay and by the challenge's own utility: every leak gap survives, $+8.8$ to $+12.8$ per-patient AUROC points and $+0.12$ to $+0.39$ utility, and the case-study pair moves from 0.775 to 0.885 per patient and from 0.33 to 0.63 utility.

What this says about the literature. Bidirectional and unmasked encoders are routinely applied to this benchmark (Das et al., 2025), usually with a fixed prediction window, which excludes the future by construction; per timestep only the architecture excludes it. We do not claim that any published

number is this artefact, only that a per-timestep number far above the causal column on this label is worth the test above.

7 CHECKLIST

For any per-timestep result:

1. **Replay every prediction on the truncated prefix** (FPS, §4.1). Any non-zero change is a causality break; zero means no leak detected, not causal. Add the value perturbations to learn what is read; never rely on shuffle alone.
2. **Compute the label-only score.** If the label is monotone, A_1 costs one line and shows what one oracle bit about the onset is worth on this label. It is a reference, not a bound: a causal result near it should say which inputs carry that information.
3. **Ablate the context to observation indicators.** Report the share of the gain recovered without any measured value: the measurement process, legitimate only if causal (§8.1, Appendix C.2).
4. **State the gate's degrees of freedom.** If $C{=}1$ the gate is the identity and that cell is vacuous; top-$m = C$ removes only sparsity, not context-dependence (§8.2).
5. **Diff the configuration files** of baseline and proposed model, and report conditioning parameters per training sequence with any gain (§8.3).

8 CASE STUDY: ANATOMY OF ONE ROUTED ADDITIVE MODEL

The remainder of the empirical part returns to the model that led us here (Anonymous, 2026). Its value as a case study is that we had its complete training scripts and configuration files, which is what makes the dissection here and in Appendix C possible.

What the artefact is. A $2{\times}2$ decomposition (Appendix C.1) settles it. Holding causality fixed, four experts buy -0.27 over GATSM (0.788 vs. 0.791); holding the mixture at zero, a non-causal rank-1 gate on the same backbone buys $+13.88$ (0.929). The reported $+14.25$ (0.933; $+14.7$ over twenty splits, Table 1) is non-causality with routing along for the ride—*the artefact*. Ablating the router's input (Appendix C.2) shows what the channel carries: a router that sees only the observation indicators—which variables were recorded, and when—recovers about half of it ($+6.50$ at five seeds, $+7.10$ at twenty); a constant gate, 3%. The model reads the measurement process, and the information enters through the gate rather than the additive path, where an attribution would have shown it.

8.1 THE PART THAT WAS REAL: INFORMATIVE MISSINGNESS, CAUSALLY

The indicator-only router of Appendix C.2 (about half the artefact) was non-causal, so it mixes two things: which labs were ordered, and when, tracks clinical suspicion; and the whole-record pattern also encodes when the record turns positive. Only the first is available at prediction time. We separate them with one change: the router reads the *running* mean of the indicators over valid timesteps $\leq k$. The additive path is untouched and the model passes FPS.

router context	n	causal	AUROC	vs. baseline†	wins
— (GATSM)	20	yes	0.7866 ± 0.0096	—	—
running-mean basis (App. C.1)	5	yes	0.7879 ± 0.0083	-0.27	1/5
running-mean basis, no pretraining	20	yes	0.7801 ± 0.0089	-0.65	5/20
indicators as *inputs*, no router	20	yes	0.8002 ± 0.0095	$+1.36$	19/20
causal indicator router (ours)	**20**	**yes**	0.8037 ± 0.0086	$+1.71$	**19/20**
whole-record indicators (App. C.2)	20	no	0.8576 ± 0.0062	$+7.10$	20/20
whole-record basis (as reported)	20	no	0.9338 ± 0.0059	$+14.72$	20/20

†paired over shared seeds, so a delta need not equal the difference of means.

The causal indicator router beats the additive baseline on 19 of 20 splits ($+1.7$, $p < 10^{-4}$; Table 8), per patient ($+2.0$) and by utility ($+0.02$; Table 2)—a real but small gain, and we say so. Most of

it needs no router: the same indicators given to the additive baseline as input columns buy $+1.36$ ($19/20$, $p < 10^{-4}$), and routing adds $+0.35$ over that control ($13/20$, $p = 0.17$; 95% interval -0.13 to $+0.83$), which twenty seeds do not resolve. The gain is the measurement process read causally, and what the router reads matters more than that it routes: on matched settings, indicators beat the basis by $+2.36$ on all twenty seeds ($p < 10^{-4}$). The remaining $+5.4$, and the whole of the reported $+14$, are the future. The measurement process is thus worth a little without breaking the additive read-out: no measured value enters the gate, so for a fixed input the logit still decomposes exactly into per-feature terms. A fair comparison of the routed model as designed gives -0.3; the $+1.7$ is what a causal redesign of its one informative input recovers.

8.2 Constraining the gate does not fix it

The obvious remedy is to narrow the channel. We vary two axes: soft versus hard (one-hot top-1, straight-through) and per-feature versus one decision shared across features. Routing is per-feature by default, so a hard gate still emits $F \log_2 C$ bits—80 at F=40, C=4—and only sharing reduces it to $\log_2 C = 2$. Cells: twenty seeds, no pretraining, baseline 0.7866.

gate	channel capacity	AUROC	artefact
per-feature, soft top-m	unbounded	0.9307 ± 0.0049	$+14.41$
per-feature, hard top-1	$F \log_2 C = 80$ bits	0.9294 ± 0.0050	$+14.28$
shared, soft top-m	one continuous decision	0.9384 ± 0.0094	$+15.18$
shared, hard top-1	$\log_2 C = 2$ bits	0.8786 ± 0.0140	$+9.20$

Neither restriction alone reduces the artefact—sharing alone raises it $+0.77$ (Figure 3, Appendix D). Both together remove 36%, leaving 64% of the artefact intact (0/20 seeds reach the soft gate, $p < 10^{-4}$). The reason is arithmetic: a binary decision needs one bit, and any mixture with $C \geq 2$ has at least one. In this gate family access, not bandwidth, binds; the remedy is to make Equation (1) causal.

8.3 Correcting the benchmark

The gate can be provably inert. With C=1 the softmax over one expert is identically 1, so the router cannot influence the output. Four of the ten reported rows use C=1 (Table 5, Appendix E); there the routed model *is* the baseline, and the reported differences (mean $|\Delta| = 3.61$, of which 1.65 survives matched augmentation) are augmentation and seed noise.

An augmentation asymmetry. The routed configuration applies time-warp augmentation (and on two datasets label smoothing) that the baseline does not; matching them changes the reported deltas substantially (Table 6).

8.4 What survives, and what tracks it

Causal context, matched augmentation and a rank-8 (low-rank) router on the six datasets whose gate is active leave, at twenty seeds, gains only on the two air-quality datasets ($+6.4$, $+3.3$; $20/20$, $p < 10^{-4}$), which share one source (Table 6, Appendix F). The gain falls with router parameters per training sequence ($r = -0.62$, $n = 6$, $p = 0.19$), not with a pre-hoc estimate of context-dependence ($r = -0.02$; Appendices H, K).

9 Limitations

One benchmark, with a monotone label that turns on six hours before Sepsis-3 onset, underlies the audit and both label-only scores. The four families are one-line switches rather than tuned models, though twelve width and depth settings keep the causal column within 0.785–0.807 (Table 9); only the recurrent pair is not parameter-matched (Appendix L). FPS detects a path, not its magnitude, and only at the inputs and cut it is run on. On the case study, the two convincing corrected gains share one air-quality source, six active-router datasets make the parameter-cost correlation suggestive, the synthetic pair also changes the local signal, the causal fix is one design, per-patient gaps vary across splits, and the audited model is our own concurrent submission (Ethics statement).

AI USE STATEMENT

We used an LLM-based assistant extensively in this work. Beyond drafting and editing text, it contributed to items ICLR requires us to disclose: it helped develop the leakage-class framing and the label-only score (Remark 1), including the statement and the proof sketch in its footnote; it proposed and refined hypotheses and the design of the matched-pair audit, the router-input ablation, the gate ladder and the input-recoverable and learned-channel analyses; it implemented the future-perturbation diagnostic, the matched-pair baselines and the experiment drivers; it produced figures and tables from logged results; it assisted with literature search; it drafted interpretations of results; and it produced simulated reviews of drafts whose points we then addressed. It generated the synthetic sequence pair of Appendix G. We did not use it to clean or relabel any real dataset, and the remaining disclosure categories (translation; use on human-subject data beyond the public de-identified benchmarks) do not apply. The research question, the decision to audit the case-study model and to report the negative result, the selection of experiments to run, and the final claims are the authors' own. We reviewed all AI-assisted work: every table is regenerated from run logs by the scripts in the supplementary material, except the marker coordinates of Figure 2, which are transcribed from the logs named beside them, and the stored result files of Appendices I and J; the diagnostic is covered by unit tests; the proof was checked by the authors; and every reference was checked against Crossref, OpenAlex, the arXiv API or the publisher's own record, with the assistant issuing the queries and the authors reading the results. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The anonymised code, configuration files, and the driver scripts for every table and figure are provided as supplementary material. The future-perturbation score is a single script that needs only `forward(x, t)` (Appendix A); the matched-pair baselines, the label-only score, and the corrected benchmark are each one shell script over fixed seeds—twenty where a table says so, five otherwise—with the splitting rule of §3. Hyperparameters, hardware and compute are listed in Appendix L. PhysioNet2019 and the UCR/UEA/Monash series are public; no data were collected.

ETHICS STATEMENT

PhysioNet2019 is a public, de-identified clinical dataset used under its published license. This work reports a failure mode of a class of early-warning models and a label-only reference for what such a leak can be worth; we see no risk of harm from its publication, and the intended effect is to reduce the chance that non-causal models are deployed on the strength of leaked per-timestep scores.

Relationship to the audited model. The routed model of Anonymous (2026) is our own concurrent submission, cited in the third person to preserve double-blind review. That is why we hold its complete configuration and training scripts and can release them; no third party's unpublished work was used. The two submissions share no claim: that manuscript proposes the architecture; this one names the leakage form and claims the detector benchmark and diagnostic perturbations, the label-only scores and the audit, and reports the architecture's headline number as an artefact.

REFERENCES

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 4699–4711, 2021.

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, 2018. doi: 10.1136/bmj.k1479.

Anonymous. Neural additive experts for interpretable time series, 2026. Manuscript under review; anonymised implementation provided as supplementary material.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271v2*, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23049–23062, 2022.

Pronaya Prosun Das, Lena Wiese, Marcel Mast, Julia Böhnke, Antje Wulff, Michael Marschollek, Louisa Bode, Henning Rathert, Thomas Jack, Sven Schamer, Philipp Beerbaum, Nicole Rübsamen, André Karch, Christian Groszweski-Anders, Andreas Haller, and Torsten Frank. An attention-based bidirectional LSTM-CNN architecture for the early prediction of sepsis. *International Journal of Data Science and Analytics*, 20(3):1841–1855, 2025. doi: 10.1007/s41060-024-00568-z.

Lucas M. Fleuren, Thomas L. T. Klausch, Charlotte L. Zwager, Linda J. Schoonmade, Tingjie Guo, Luca F. Roggeveen, Eleonora L. Swart, Armand R. J. Girbes, Patrick Thoral, Ari Ercole, Mark Hoogendoorn, and Paul W. G. Elbers. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, 46(3):383–400, 2020. doi: 10.1007/s00134-019-05872-y.

Xavier Fonseca. Look-ahead-freedom as temporal non-interference: A verifiable correctness property for backtesting and agentic trading pipelines. *arXiv preprint arXiv:2607.04958v1*, 2026.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6):602–610, 2005. doi: 10.1016/j.neunet.2005.06.042.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. doi: 10.1111/j.2517-6161.1993.tb01939.x.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18):eadk3452, 2024. doi: 10.1126/sciadv.adk3452.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4):1–21, 2012. doi: 10.1145/2382577.2382579.

Minkyu Kim, Suan Lee, and Jinho Kim. Transparent networks for multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37510–37518, 2026. doi: 10.1609/aaai.v40i44.41084.

Michael Moor, Bastian Rieck, Max Horn, Catherine R. Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Frontiers in Medicine*, 8:607952, 2021. doi: 10.3389/fmed.2021.607952.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 8414–8426, 2022.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology challenge 2019. *Critical Care Medicine*, 48(2):210–217, 2020. doi: 10.1097/CCM.0000000000004145.

Ronald Richman and Mario V. Wüthrich. LocalGLMnet: interpretable deep learning for tabular data. *Scandinavian Actuarial Journal*, 2023(1):71–95, 2023. doi: 10.1080/03461238.2022.2081816.

Giacinto Paolo Saggese and Paul Smith. Causify DataFlow: A framework for high-performance machine learning stream computing. *arXiv preprint arXiv:2512.23977v1*, 2025.

Ikaro Silva, George Moody, Daniel J. Scott, Leo A. Celi, and Roger G. Mark. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology challenge 2012. In *Computing in Cardiology*, volume 39, pp. 245–248, 2012.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.05109v4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

Guangzhi Xiong, Sanchit Sinha, and Aidong Zhang. Neural additive experts: Context-gated experts for controllable model additivity. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300 of *Proceedings of Machine Learning Research*, pp. 775–783, 2026. arXiv:2602.10585.

A THE CAUSALITY DIAGNOSTIC IN FULL

The check requires no training and no access to labels. On randomly initialised weights the routed model already reports 9.1×10^{-3} (unnormalised), the baseline exactly 0.

```
import torch
B, T, F, CUT, TOL = 2, 12, 5, 6, 1e-6
t, x = torch.tensor([T-1, T-1]), torch.randn(B, T, F)
def fps(model, x, t):                          # Equation (2)
    model = model.double().eval(); x = x.double()   # float64
    with torch.no_grad():
        y1 = model(x, t); sd = y1.std(unbiased=False).clamp_min(1e-12)
        x_shift = x.clone(); x_shift[:, CUT+1:] += 5.0     # shift future
        x_shuf  = x.clone()
        for b in range(B):                          # shuffle future
            x_shuf[b, CUT+1:] = x[b, CUT+1:][torch.randperm(T-CUT-1)]
        x_scale = x.clone(); x_scale[:, CUT+1:] *= 3.0     # scale future
        x_warp = x.clone(); f = x_warp[:, CUT+1:]      # warp future
        x_warp[:, CUT+1:] = f.sign() * f.abs().pow(1.5)
        x_trunc = x[:, :CUT+1].clone()                 # prefix replay
        t_trunc = t.clamp(max=CUT)                     # (crop)
        out = {}
        cases = {"shift": (x_shift, t), "shuffle": (x_shuf, t),
                 "scale": (x_scale, t), "warp": (x_warp, t),
                 "trunc": (x_trunc, t_trunc)}
        for name, (xp, tp) in cases.items():
            y2 = model(xp, tp)[:, :CUT+1]
            d = (y1[:, :CUT+1] - y2).abs().mean(0).max()
```

```
            out[name] = float(d / sd)
    return max(out.values()), out
for name, model in [("baseline", baseline), ("routed", routed)]:
    score, parts = fps(model, x, t)
    print(name, parts, "LEAK" if score > TOL else "no leak detected")
```

The released `causalcheck.py` adds batching over held-out data, a fixed permutation seed, and scores in float64. The replay recomputes on a shorter tensor, which changes the order of float32 sums: with TF32 on, causal LSTMs showed a spurious $\sim 10^{-4}$, and even in full float32 causal attention models showed up to 4×10^{-7}, on CPU as well. In float64 every one of the 125 causal checkpoints of Table 1 scores exactly 0 under all five perturbations, and the weakest leak scores 0.43. One checkpoint, twenty held-out batches and all five perturbations take 25–40 s on one A6000, mostly data loading, with about 3 GB of host memory. Unit tests cover a pooled context that shuffle misses, statistics that only scale or warp expose, a model that reads only the record length, one that reads the tensor's extent, and one with a large output offset. Table 3 compares detectors on synthetic leaks. Figure 1 uses trained checkpoints and plots the unnormalised mean absolute change per timestep

Table 3: Which detector catches which leak (14 synthetic prototypes, each with a causal twin that reads the same statistic over steps $\leq k$; cut $= 15$, flag $> 10^{-6}$; `bench_detectors.py`). Prefix replay is the two-run test; the value perturbations add diagnosis, not coverage.

leak read by the model	shuffle	shift	four value	zero-pad trunc.	prefix replay	resample	FPS
whole-sequence mean	—	✓	✓	✓	✓	✓	✓
whole-sequence variance	—	✓	✓	✓	✓	✓	✓
whole-sequence skewness	—	✓	✓	✓	✓	✓	✓
fraction of positive values	—	✓	✓	✓	✓	✓	✓
variance of the future segment	—	✓	✓	✓	✓	✓	✓
skewness of the future segment	—	✓	✓	✓	✓	✓	✓
variance of a window beyond the cut	✓	—	✓	✓	✓	✓	✓
skewness of a window beyond the cut	✓	—	✓	✓	✓	✓	✓
last valid index t	—	—	—	✓	✓	—	✓
tensor extent	—	—	—	—	✓	—	✓
batch-wide mean	—	✓	✓	✓	✓	✓	✓
bidirectional LSTM	✓	✓	✓	✓	✓	✓	✓
attention without mask	—	✓	✓	✓	✓	✓	✓
symmetric convolution	✓	✓	✓	✓	✓	✓	✓
caught	4/14	10/14	12/14	13/14	14/14	12/14	14/14
false positives	0	0	0	0	0	0	0
forward calls	2	2	5	2	2	2	6

(the quantity inside the $\max$ above, before division by $\mathrm{sd}\, f(x)$).

B LABEL-ONLY AND INPUT-RECOVERABLE SCORES: PROCEDURE

Both scores use the 60/20/20 patient split of every trained model. Per seed there are 24,202 training, 8,067 validation and 8,067 test patients; 1,718–1,777 training patients turn septic, with 16.3–16.9k positive of about 930k training timesteps and 5.3–5.9k positive of about 310k test timesteps.

1. *Cells.* For A_b a patient's cell is $c(\tau)$: one cell for never-septic and $2^b - 1$ quantile bins of the training fold's septic onsets ($b = 1$: septic or not). For $\hat{A}_b$ the cell is predicted: a probe maps the whole-record summary $g(x)$—each variable's mean over the stay, the fraction of hours it was recorded, and the stay length—to $\Pr[\text{septic}]$; $b = 1$ thresholds the probe at the training prevalence and $b \geq 2$ takes 2^b quantile bins of its score. Training cells are cross-fitted (five folds), test cells come from the probe fitted on the whole training fold, and no label enters g. For the split by source, `ICULOS` and `HospAdmTime` form the time columns: `ICULOS` counts hours since ICU admission, so its whole-stay mean tracks the stay length, while its first value, the ICU offset, is known when the record starts.
2. *Posterior, training fold only.* $s(c, k)$ is the fraction of training patients in cell c whose stay reaches k and who have $y_k = 1$; the denominator is that risk set, so shorter stays drop out as k grows. A pair (c, k) unseen in training scores the training septic prevalence.
3. *Score.* Every test pair (patient, k) inside the stay receives $s(c, k)$ for its test cell, and AUROC is pooled over all test pairs, exactly as for the trained models.

The oracle cell uses the test onset by design—A_b is the score of a channel that carries it—but no posterior is ever estimated on test labels. $A_1^\star$ and $A_2^\star$ choose threshold partitions on the validation fold (candidates: training-onset quantiles in steps of 2.5% for $b = 1$ and 5% for $b = 2$, plus the side the never-septic cell joins) and report the selected partition's test AUROC. Scripts: `label_smuggling_ceiling.py`, `ceiling_astar.py`, `ceiling_recover.py`.

Table 4: Oracle versus input-recoverable label-only scores on PhysioNet2019 (five split seeds; per-timestep AUROC). Rows below the rule replace the oracle cell $c(\tau)$ by a cell predicted from a whole-record summary $g(x)$, or (last block) from the part of a trained non-causal model's first output that comes from the future; "probe" is the probe's AUROC for septic-at-any-time on the test fold. $A_0 = 0.703 \pm 0.010$ (time index only).

cell from	probe	$b = 1$	$b = 2$	$b = 3$	$b = 4$
oracle cell $c(\tau)$ (A_b)	—	0.957 ± 0.002	0.987 ± 0.001	0.995 ± 0.000	0.998 ± 0.000
clinical values, no time/static, boosted	0.831 ± 0.012	0.757 ± 0.006	0.810 ± 0.007	0.808 ± 0.007	0.794 ± 0.006
recording rates, boosted	0.877 ± 0.005	0.797 ± 0.007	0.837 ± 0.006	0.840 ± 0.004	0.835 ± 0.004
clinical values + rates, boosted	0.896 ± 0.005	0.806 ± 0.006	0.848 ± 0.004	0.857 ± 0.003	0.852 ± 0.006
ICU offset at start (causal), boosted	0.504 ± 0.013	0.701 ± 0.010	0.691 ± 0.012	0.677 ± 0.016	0.665 ± 0.015
time columns, boosted	0.848 ± 0.005	0.801 ± 0.007	0.848 ± 0.005	0.858 ± 0.005	0.865 ± 0.006
stay length, boosted	0.829 ± 0.009	0.811 ± 0.008	0.862 ± 0.007	0.883 ± 0.005	0.889 ± 0.007
stay length, logistic	0.528 ± 0.008	0.727 ± 0.010	0.877 ± 0.008	0.906 ± 0.005	0.917 ± 0.005
all summaries (every column), length, boosted	0.938 ± 0.004	0.858 ± 0.004	0.872 ± 0.007	0.895 ± 0.005	0.901 ± 0.003
trained channel: future-derived shift of the first output					
routed model as reported	0.575 ± 0.008	0.852 ± 0.008	0.868 ± 0.008	0.889 ± 0.010	0.891 ± 0.011
LSTM + sequence mean	0.517 ± 0.008	0.725 ± 0.008	0.873 ± 0.006	0.893 ± 0.005	0.900 ± 0.005
BiLSTM	0.661 ± 0.069	0.821 ± 0.009	0.838 ± 0.008	0.841 ± 0.008	0.827 ± 0.011
unmasked Transformer	0.625 ± 0.020	0.830 ± 0.008	0.825 ± 0.016	0.832 ± 0.016	0.821 ± 0.017
symmetric-padding TCN	0.616 ± 0.015	0.720 ± 0.006	0.724 ± 0.015	0.710 ± 0.016	0.682 ± 0.015

C Case study: decomposition and channel ablation

C.1 Decomposition

model	causal	experts	AUROC	vs. baseline
GATSM	yes	1	0.7906 ± 0.0055	—
routed + causal router (running-mean basis)	yes	4	0.7879 ± 0.0083	-0.27
routed, as reported	no	4	0.9331 ± 0.0056	$+14.25$
GATSM + non-causal rank-1 gate	no	0	0.9294 ± 0.0098	$+13.88$

Two readings, and they agree (§8): four experts under causality buy -0.27; non-causal access with the mixture held at zero buys $+13.88$. The mixture is alive in both models (Appendix I); a synthetic pair reproduces the artefact (Appendix G).

C.2 What the channel carries

If the router's advantage is access to the sequence, what in the sequence? We ablate the router's input at fixed $C{=}4$, leaving the additive path untouched. All cells are the five seeds of Appendix C.1, so deltas are against its GATSM (0.7906), not the twenty-seed 0.7866 of §8.1; the shares change by under a point.

router sees	AUROC	vs. baseline	share
nothing (constant gate)	0.7944 ± 0.0068	$+0.38$	3%
observation indicators only	0.8556 ± 0.0081	$+6.50$	**46%**
pooled basis (as reported; re-run)	0.9324 ± 0.0043	$+14.18$	100%

The pooled row is re-trained inside this sweep, hence 0.9324 against 0.9331 in the previous table. The middle row matters. That router never sees a measured value—only a binary mask of which variables were recorded at which timesteps. It recovers nearly half the effect. The model is reading the measurement process: which labs a clinician ordered, and when. This is invisible to an additive attribution, because the additive path never sees the mask; the information enters through the gate.

D GATE LADDER: FIGURE AND PROTOCOL

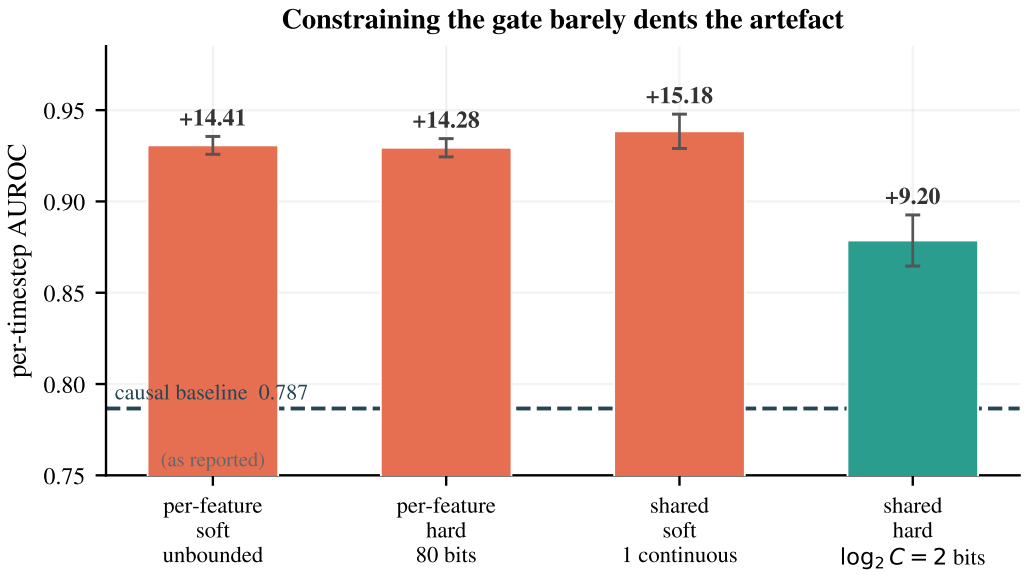

Figure 3: The artefact under four gate configurations on PhysioNet2019. Neither restriction alone reduces it; both together remove 36% (twenty seeds).

All cells are PhysioNet2019, non-causal router, $C{=}4$, no pretraining, twenty seeds, per-timestep AUROC. The twenty-seed causal baseline is 0.7866 ± 0.0096. A label-driven oracle gate reaches 1.000 even at 2 bits (PhysioNet2012; per-feature cell $n = 1$, shared cell $n = 5$), so a capacity bound of this kind is not measurable on a binary task.

E THE COMPLETE CORRECTED BENCHMARK

Table 5 gives all ten datasets. *Reported* is the delta as it appears in Anonymous (2026); *corrected* removes the augmentation asymmetry (and, for PhysioNet2019, the causality break). C is the expert count selected on validation in the original work; rows with $C{=}1$ have a provably inert gate, so the routed model and the baseline are the same function and the reported delta measures configuration differences alone. $\Delta p/N$ is router parameters per training sequence.

Table 5: All ten benchmarks after correction. $\hat{B}$ is the pre-hoc context-dependence estimate of Appendix H (axis-searched, null-adjusted).

dataset	C	gate	reported Δ	corrected Δ	$\hat{B}$	$\Delta p/N$
NATOPS	1	inert	+3.34	+0.56	0.882	266.7
Heartbeat	1	inert	−2.65	−2.65	1.128	1506.5
AppliancesEnergy	1	inert	−8.06	−3.35	0.371	694.0
LiveFuelMoistureContent	1	inert	−0.37	−0.04	2.240	1.6
LSST	2	active	+2.83	+1.61	0.942	2.4
AustraliaRainfall	2	active	−0.07	−0.14	1.312	0.0
BeijingPM10Quality	4	active	+4.72	+3.83	0.686	3.2
BeijingPM25Quality	4	active	+2.93	+2.27	0.360	3.2
PhysioNet2012	4	active	−1.13	−1.13	0.469	155.9
PhysioNet2019	4	active	+14.25	−0.27	0.342	26.4

Across the four inert rows the mean absolute as-reported delta is 3.61 points; 1.65 remains after matching the augmentation, and since the two models are the same function that remainder is seed noise.

F WHAT SURVIVES: FULL NUMBERS

The learned context is real. Fixing C=4 and router rank 8, we vary only what drives the gate.

gate	PhysioNet2012	BeijingPM10	LSST
baseline (no gate)	0.8412	0.5749	0.5673
constant	0.8359	0.5728	0.5639
frozen random	0.8338	0.6053	0.5756
learned context	**0.8485**	**0.6291**	**0.5838**
learned − random	+1.47 (5/5)	+2.38 (5/5)	+0.82 (4/5, p=0.562)

Learned context beats both controls on all three datasets, decisively on two and weakly on LSST. Random instance-dependent variation is not a substitute: it is *worse* than a constant gate on PhysioNet2012 and better on the other two, with no consistent sign.

. . . and it is mostly spent on its own machinery. Note where the gain goes on PhysioNet2012. Adding experts costs 0.53; routing them at random costs 0.74; learning the routing repays that and lands 0.73 above baseline, which is not significant at five seeds (and −0.06 at twenty; below). **Conditional routing buys a genuine but small amount of signal, most of which is spent paying for the apparatus that carries it.**

Router size versus context-dependence: an underpowered comparison. The full router costs F^2KC parameters (K basis functions per feature); a rank-r factorisation costs $rF(K + C)$. On PhysioNet2012, five paired seeds: rank-8 versus full router +1.86 (5/5, $p = 0.062$); full router versus baseline −1.13 (1/5); rank-8 versus baseline +0.73 (3/5); at twenty seeds rank-8 versus baseline is −0.06 (9/20), so the PhysioNet2012 gain does not survive. Same direction with a causal router on PhysioNet2019 (+0.72, 4/5). Across the six datasets with an active router, the corrected (pre-rank-8) delta falls as router parameters per training sequence rise ($r = -0.62$, $p = 0.19$; Spearman −0.46; Pearson −0.47 without PhysioNet2012, whose $\Delta p/N = 155.9$ is the leverage point) and is not predicted by a pre-hoc estimate of context-dependence ($r = -0.02$; Appendix K).

Table 6: Applying all three corrections—causal context, matched augmentation, and a rank-8 router—on every dataset whose gate is actually active. AUROC/Acc/R^2 points. The rank-8 column is at twenty seeds, paired against twenty-seed baselines (Appendix L); the other columns are at five.

dataset	reported Δ	corrected Δ	**corrected + rank-8 Δ**
LSST	+2.83	+1.61	−0.96
AustraliaRainfall	−0.07	−0.14	+0.12
BeijingPM10Quality	+4.72	+3.83	+6.42
BeijingPM25Quality	+2.93	+2.27	+3.30
PhysioNet2012	−1.13	−1.13	−0.06
PhysioNet2019	+14.25	−0.27	+0.02 (causal)
mean (mixed metrics; descriptive only)	+3.92	+1.03	+1.47

What a corrected, correctly-sized model delivers. The reported table is erratic: it swings from −8.1 to +14.2 and half its rows are vacuous or confounded. A corrected and correctly-sized routed model is *not* better on average than the reported one claims to be—and at twenty seeds only the two air-quality gains survive (+6.4, +3.3; 20/20); the other four—LSST, AustraliaRainfall, PhysioNet2012 and the causal PhysioNet2019 cell—are within noise, −1.0 to +0.1. The caveat is that the two individually convincing gains come from one air-quality source and are not independent.

The practical reading: conditional routing in additive time-series models is a small, real, and fragile effect. It is worth having, it is not worth the parameter budget the design as reported spends on it, and it is very easy to confuse with a causality break.

G THE CONTROLLED SYNTHETIC PAIR

PhysioNet2019 is the suite's only many-to-many task, so we construct a matched synthetic pair with identical length distributions, class balance and onset structure, designed to differ in whether a sequence-level statistic reveals the class (they also differ in local signal; see below) (AUROC of the sequence mean $\rightarrow$ class: 0.77 for LEAK, 0.50 for NO-LEAK). The non-causal router gains $+3.18$ AUROC on LEAK (5/5 seeds) and $+0.51$ on NO-LEAK (4/5, p=0.125); the causal router gains $+0.14$ and $+0.28$. The artefact appears at full size only when sequence-level statistics carry class information (Figure 4, Table 7).

Each sequence has length $T \sim \mathrm{Unif}(30, 80)$, twelve AR(1) features with $\rho = 0.7$, and a positive-class rate of 0.31. A positive sequence turns its label on at an onset $\tau \sim \mathrm{Unif}(8, T-6)$. The variants differ only in the signal applied to the first three features:

- LEAK: $+\delta$ for all $t \geq \tau$ (a persistent level shift, so the sequence mean carries the class);
- NO-LEAK: $+\delta$ on $[\tau, \tau+w)$ then $-\delta$ on $[\tau+w, \tau+2w)$ for positives and the reverse for negatives, $w = 5$ (zero-mean over the sequence, up to truncation of the compensating step for onsets later than $T-2w$).

The two variants therefore also differ in local signal structure, which is why their causal baselines differ (0.826 vs. 0.714 at $\delta = 0.6$); every comparison is within-variant. With $\delta = 1.6$ the baseline saturates near 0.97 on LEAK; the low-SNR variant used in the main text sets $\delta = 0.6$. The resulting AUROC of the sequence mean of the three signal features as a class predictor is 0.938 / 0.766 for LEAK at the two amplitudes and 0.504 / 0.496 for NO-LEAK (averaging all twelve features gives 0.826 / 0.646 for LEAK).

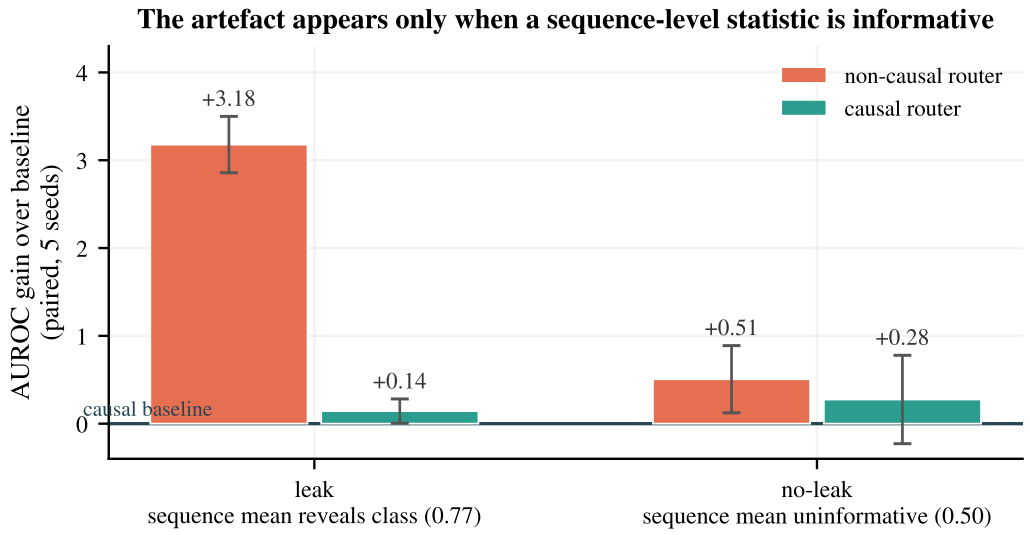

Figure 4: Paired AUROC gain over the causal baseline on the controlled pair (5 seeds, error bars are the standard deviation of the paired differences). The non-causal router gains only when a sequence-level statistic is informative.

Table 7: The controlled pair: per-timestep AUROC, 5 seeds. The main text uses the low-amplitude variant ($\delta = 0.6$); at $\delta = 1.6$ the non-causal router gains $+1.51$ where the sequence mean carries the class and *loses* 2.41 where it does not ($0/5$), consistent with the gain coming from the sequence-level channel; because the variants also differ in local signal, the pair illustrates the mechanism rather than isolating it.

variant	model	AUROC	vs. baseline	wins
LEAK	GATSM	0.8261 ± 0.0096	—	—
LEAK	**non-causal router**	0.8579 ± 0.0109	$+3.18$	**5/5**
LEAK	causal router	0.8276 ± 0.0088	$+0.14$	5/5
NO-LEAK	GATSM	0.7135 ± 0.0072	—	—
NO-LEAK	non-causal router	0.7185 ± 0.0093	$+0.51$	4/5 (p=0.125)
NO-LEAK	causal router	0.7162 ± 0.0077	$+0.28$	3/5 (p=0.438)
high-amplitude variant, $\delta = 1.6$				
LEAK	GATSM	0.9703 ± 0.0009	—	—
LEAK	non-causal router	0.9855 ± 0.0004	$+1.51$	5/5
LEAK	causal router	0.9675 ± 0.0024	-0.29	1/5
NO-LEAK	GATSM	0.8141 ± 0.0041	—	—
NO-LEAK	non-causal router	0.7901 ± 0.0120	-2.41	0/5
NO-LEAK	causal router	0.8167 ± 0.0182	$+0.26$	2/5

H PRE-HOC CONTEXT-DEPENDENCE ESTIMATOR

We estimate the context-dependence energy $B = \mathbb{E}_z \sum_f \|g_f^*(\cdot \mid z) - \bar{g}_f(\cdot)\|^2$ without training any deep model. Sequences are described by per-feature mean and standard deviation over time, standardised, and split into four equal-sized strata along a principal direction. An independent additive spline model (cubic B-spline bases per feature, ridge-penalised, fit jointly) is estimated per stratum on the training split only, and shape functions are evaluated on a common grid of inner quantiles and centred. Then

$$\hat{B} = \frac{\sum_f \mathbb{E}_v \left[\mathrm{Var}_s \, g_{f,s}(v)\right]}{\sum_f \mathbb{E}_v \left[\bar{g}_f(v)^2\right]}.$$

Because a router may learn any partition, we search the five leading principal directions and take the maximum. Small strata produce dispersion by noise alone, so we subtract a permutation null in which stratum labels are shuffled at the sequence level; the null uses the same selection rule (maximum over five random partitions), otherwise the search inflates $\hat{B}$. Reported values are $\hat{B} - \hat{B}_{\text{null}}$.

Neither the single-axis nor the axis-searched variant predicts the corrected deltas across the six active-router datasets ($r = +0.23$ and -0.02). The estimator is cheap and we report it as a negative result rather than a tool.

I ORDER PARAMETERS OF THE MIXTURE

Measured on the PhysioNet2019 test split, three seeds. $C_{\text{eff}} = \exp H$ where H is the mean entropy of the routing distribution; *spread* is the mean standard deviation of routing weights across instances; *divergence* is $\|\theta_c - \bar{\theta}\|_F / \|\bar{\theta}\|_F$ averaged over experts. One might object that the routed model simply failed to use its experts. It did use them: the reported model has effective expert count 1.41, routing spread 0.194 and expert parameter divergence 1.18; the causal model 1.23, 0.180 and 0.76. The machinery is comparably alive in both; only one of them gains.

model	C_{eff}	top-1 share	spread	divergence
as reported (non-causal)	1.41	0.84	0.194	1.18
causal router	1.23	0.92	0.180	0.76

The experts differentiate and routing varies across instances in both models. The mixture has not collapsed; it simply does not explain the performance difference.

J ROUTING SHIFTS AND SEPSIS ONSET

With causal routing the weights become per-timestep, so we can ask whether the point at which routing changes most sharply relates to clinical onset. We score $s_t = \frac{1}{2}\,\mathbb{E}_f \|w_{t,f} - w_{t-1,f}\|_1$ and z-score it against the population at the same t, because a running-mean context is unstable early in a stay. Patient score is $\max_t z_t$ after a six-step burn-in. Measured on the causal running-mean basis router of Appendix C.1 (`-causal-C4` checkpoints) and on stays of at least twelve steps ($n \approx 7{,}900$), so the 0.792 risk-trajectory AUROC is not directly comparable with any row of Table 2.

Four seeds (1–4) were scored. Discrimination is real but weak: patient-level AUROC 0.642 against a permutation null of 0.505 (4/4 seeds above null), far below the model's own risk trajectory (0.792, 0/4). Timing is refuted: the fraction of septic patients whose routing shift precedes label onset is 0.656 against a null of 0.700 (0/4 seeds above null). Routing reacts to onset rather than anticipating it.

K HYPOTHESES WE TESTED AND REJECTED

A pre-hoc predictor of achievable gain. The context-dependence estimator of Appendix H predicts neither the corrected deltas nor their sign; numbers there. We report this as unsupported rather than refuted: with six datasets and deltas whose seed noise is on the order of one AUROC point, the test is underpowered.

Routing shifts as an onset signal. Discrimination is weak but real; the timing claim is refuted. Definitions and numbers in Appendix J.

L PROTOCOL, HYPERPARAMETERS AND COMPUTE

Splits are 60/20/20 by sequence, re-drawn per seed with the seed as the random state, so the split itself varies across seeds; preprocessing (forward fill, then mean/mode imputation, ordinal encoding, standardisation) is fit on the training split only. Training uses AdamW with early stopping on validation loss (patience 20) and no epoch cap. Architecture hyperparameters are taken unchanged from the audited work; where a comparison required matching them, the two configurations are identical apart from the factor under study, and we state which. The audited model includes a 20-epoch masked-reconstruction pretraining stage. It is on in the as-reported and running-mean-basis rows of Appendix C.1 and off in every other ablation cell (Appendix C.2, the new row of §8.1 and all of §8.2, whose table is reported at twenty seeds without pretraining). The running-mean basis router is reported both ways, with pretraining (-0.27, five seeds) and without (-0.65, twenty seeds), so the §8.1 contrast is matched on that axis. The routed model also adds a variance penalty across expert outputs (weight 0.1) to the task loss, which shrinks the experts toward agreement; we keep it as configured in every routed cell. The rank-8 router was fixed a priori and not swept, so the corrected+rank-8 column of Table 6 is a single configuration scored on the test folds, not a tuned one. The BiLSTM member of the recurrent pair has twice the recurrent width of its causal twin (the one-line switch doubles the output dimension); the other three pairs are parameter-matched. Width does not explain the recurrent gap: causal LSTMs of width 128 and 256 stay at 0.785–0.797 (Table 9), 18 points below the BiLSTM (20 splits). Seeds fix the split, initialisation, shuffling and augmentation; deterministic CUDA kernels were not enabled, so re-runs are close but not bit-exact. All experiments ran on two RTX A6000 GPUs; the full audit is approximately 350 training runs. On PhysioNet2019 the additive baseline has 123k trainable parameters, the same model with indicator input columns 130k, the causal indicator router 141k and the routed model as reported 775k. None caches state, so deployed hourly each re-runs its prefix: at batch size 1 on one CPU thread of a shared machine (medians of repeated runs, so only indicative) the three causal variants take 6–10 ms per issued prediction at 24 hours and 44–92 ms at 192 (`bench_online_latency.py`). The router's increment over indicator inputs is not resolved (§8.1), so indicator inputs are the default we would recommend.

Departures from the PhysioNet2019 challenge protocol. We score pooled per-timestep AUROC over all (patient, timestep) pairs rather than the challenge's clinical utility score with its early-prediction window; `SepsisLabel` is used as supplied, so it turns on six hours before Sepsis-3 onset

and "onset" in this paper means the label's turn-on; only the public training sets A and B are used, re-split by patient; and the eight UCR/UEA/Monash datasets are likewise re-split 60/20/20 rather than scored on the archive partitions, so Table 5 is not comparable to published archive numbers. Per-patient AUROC and the challenge utility are in Table 2. For the routed model as reported, splits 0–4 use the shipped checkpoints and splits 5–19 fifteen trainings of its configuration file (pretraining on, C from the file); its per-patient gap over GATSM ranges from $+7.3$ to $+16.5$ across the twenty splits and its utility gap is at least $+0.26$ (an earlier single training of that configuration, outside these twenty, gave only $+0.4$ per patient), so per-patient numbers vary more across splits than pooled ones, which we flag rather than hide.

Multiplicity. About forty paired tests appear in the main text; the five-seed ones have a Wilcoxon floor of $p = 0.062$ and none of those survives a Bonferroni correction. The decisive comparisons were therefore re-run at twenty seeds (Table 8): all four leak gaps of Table 1, the artefact itself, the shared hard gate's reduction and the causal indicator router reach $p < 10^{-4}$, below the Bonferroni threshold of 0.00125 for forty tests, while the per-feature hard gate, the LSST, AustraliaRainfall and PhysioNet2012 rank-8 gains and the causal PhysioNet2019 rank-8 cell do not differ from their baselines. Sharing the gate while leaving it soft raises the artefact $+0.77$ ($p = 0.002$), above the corrected threshold. Combining the four families of Table 1 by sign alone gives $p = 3 \times 10^{-19}$ (Fisher, on four exact sign-test p-values), descriptive since the pairs share splits. The 36% gate reduction is a ratio of means whose per-seed values range from 16% to 61%.

What the seeds estimate. Each seed redraws the 60/20/20 split and the initialisation, so a paired difference over seeds estimates the mean effect of the training procedure on this fixed public cohort, with the interval in Table 8; it is not a patient-level statement about a new cohort, and timesteps are never treated as independent samples. The primary comparisons are the five leak gaps of Table 1 and the causal indicator router against GATSM; every other test in Table 8 and the text is secondary and is read against the Bonferroni threshold above, or reported as an effect size where it does not reach it.

Power. Table 8 repeats the decisive comparisons at twenty seeds (seeds 0–19; the first five are the runs of the main text), with exact paired Wilcoxon p-values.

Table 8: Twenty-seed re-runs of the decisive comparisons (per-timestep AUROC on PhysioNet2019, accuracy/R^2 elsewhere; A minus B in points with the 95% t-interval of the paired mean over seeds; exact paired Wilcoxon). P19/P12: PhysioNet2019/2012; PM10/PM25: BeijingPM10/25Quality; Rainfall: AustraliaRainfall.

comparison	n	A	B	Δ [95% CI]	wins / p
causal indicator router vs GATSM (P19)	20	0.804 ± 0.009	0.787 ± 0.010	$+1.71$ [$+1.17$, $+2.25$]	19/20, $p < 10^{-4}$
sym-pad TCN vs left-pad TCN (P19)	20	0.819 ± 0.011	0.757 ± 0.012	$+6.17$ [$+5.76$, $+6.58$]	20/20, $p < 10^{-4}$
unmasked vs masked Transformer (P19)	20	0.974 ± 0.002	0.795 ± 0.010	$+17.83$ [$+17.38$, $+18.28$]	20/20, $p < 10^{-4}$
seq-mean vs running-mean pooling (P19)	20	0.924 ± 0.005	0.781 ± 0.009	$+14.25$ [$+13.86$, $+14.64$]	20/20, $p < 10^{-4}$
rank-8 causal basis router vs GATSM (P19)	20	0.787 ± 0.010	0.787 ± 0.010	$+0.02$ [-0.53, $+0.56$]	9/20, $p = 0.927$
BiLSTM vs LSTM (P19)	20	0.974 ± 0.001	0.782 ± 0.011	$+19.23$ [$+18.71$, $+19.76$]	20/20, $p < 10^{-4}$
gate: per-feature hard vs soft (P19)	20	0.929 ± 0.005	0.931 ± 0.005	-0.13 [-0.39, $+0.14$]	9/20, $p = 0.388$
gate: shared soft vs soft (P19)	20	0.938 ± 0.009	0.931 ± 0.005	$+0.77$ [$+0.34$, $+1.21$]	16/20, $p = 0.002$
gate: shared hard vs soft (P19)	20	0.879 ± 0.014	0.931 ± 0.005	-5.21 [-5.87, -4.54]	0/20, $p < 10^{-4}$
gate: soft vs GATSM (P19)	20	0.931 ± 0.005	0.787 ± 0.010	$+14.41$ [$+13.94$, $+14.87$]	20/20, $p < 10^{-4}$
gate: shared hard vs GATSM (P19)	20	0.879 ± 0.014	0.787 ± 0.010	$+9.20$ [$+8.38$, $+10.01$]	20/20, $p < 10^{-4}$
rank-8 vs matched baseline (LSST)	20	0.553 ± 0.047	0.562 ± 0.021	-0.96 [-2.79, $+0.86$]	9/20, $p = 0.360$
rank-8 vs matched baseline (PM10)	20	0.639 ± 0.024	0.575 ± 0.025	$+6.42$ [$+5.24$, $+7.61$]	20/20, $p < 10^{-4}$
rank-8 vs matched baseline (PM25)	20	0.788 ± 0.018	0.755 ± 0.015	$+3.30$ [$+2.42$, $+4.18$]	20/20, $p < 10^{-4}$
rank-8 vs matched baseline (Rainfall)	20	0.053 ± 0.039	0.052 ± 0.038	$+0.12$ [-0.15, $+0.39$]	12/20, $p = 0.277$
rank-8 vs GATSM (P12)	20	0.834 ± 0.015	0.835 ± 0.016	-0.06 [-0.92, $+0.80$]	9/20, $p = 0.648$
routed, pooled basis, as reported vs GATSM (P19)	20	0.934 ± 0.006	0.787 ± 0.010	$+14.72$ [$+14.16$, $+15.28$]	20/20, $p < 10^{-4}$
whole-record indicator router vs GATSM (P19)	20	0.858 ± 0.006	0.787 ± 0.010	$+7.10$ [$+6.61$, $+7.59$]	20/20, $p < 10^{-4}$
running-mean basis, no pretraining vs GATSM (P19)	20	0.780 ± 0.009	0.787 ± 0.010	-0.65 [-0.98, -0.32]	5/20, $p = 7.1 \times 10^{-4}$
GATSM + indicator inputs, no router vs GATSM (P19)	20	0.800 ± 0.010	0.787 ± 0.010	$+1.36$ [$+0.95$, $+1.77$]	19/20, $p < 10^{-4}$
causal indicator router vs GATSM + indicator inputs (P19)	20	0.804 ± 0.009	0.800 ± 0.010	$+0.35$ [-0.13, $+0.83$]	13/20, $p = 0.165$
causal indicator router vs running-mean basis, no pretraining (P19)	20	0.804 ± 0.009	0.780 ± 0.009	$+2.36$ [$+1.91$, $+2.81$]	20/20, $p < 10^{-4}$

Causal-column sweep. To check that the causal column is not an artefact of one small configuration, the causal LSTM and masked Transformer were re-trained at hidden width 64, 128, 256 and depth 2, 4 on split seeds 0–2 (Table 9). No setting leaves 0.785–0.807; the best is 15 points below the one-bit label-only score and 5 below the input-recoverable one.

Table 9: Causal-column sweep on PhysioNet2019: per-timestep AUROC of the causal members of the recurrent and Transformer pairs over hidden width and depth (split seeds 0–2 per cell; the audited configuration is $h = 64$, two layers).

architecture	hidden	layers	n	AUROC
LSTM (causal)	64	2	3	0.796 ± 0.004
LSTM (causal)	64	4	3	0.796 ± 0.003
LSTM (causal)	128	2	3	0.788 ± 0.004
LSTM (causal)	128	4	3	0.785 ± 0.005
LSTM (causal)	256	2	3	0.797 ± 0.006
LSTM (causal)	256	4	3	0.793 ± 0.003
masked Transformer	64	2	3	0.807 ± 0.006
masked Transformer	64	4	3	0.803 ± 0.003
masked Transformer	128	2	3	0.800 ± 0.003
masked Transformer	128	4	3	0.785 ± 0.021
masked Transformer	256	2	3	0.798 ± 0.003
masked Transformer	256	4	3	0.790 ± 0.003

Bounded look-ahead. The output at step k of the pooled-context LSTM and of the Transformer was allowed to read steps $\leq k + L$ for $L \in \{2, 6, 12, 24\}$ hours (split seeds 0–2; per-layer window $L/2$ in the two-layer Transformer, so the effective look-ahead matches). These are audit settings, not deployable models.

Table 10: Bounded look-ahead on PhysioNet2019: per-timestep AUROC when the output at k may read steps $\leq k + L$ (split seeds 0–2). $L = 0$ and full are the causal and non-causal members of Table 1; the two-layer Transformer uses a per-layer window of $L/2$, so both families have the same effective look-ahead. The label turns on six hours before Sepsis-3 onset.

look-ahead L (hours)	0	2	6	12	24	full
LSTM + pooled context	0.794 ± 0.006	0.791 ± 0.005	0.795 ± 0.006	0.811 ± 0.004	0.901 ± 0.004	0.929 ± 0.002
Transformer	0.807 ± 0.006	0.807 ± 0.006	0.839 ± 0.003	0.935 ± 0.001	0.959 ± 0.002	0.975 ± 0.001

Every table in this paper is regenerated from the raw run logs by the scripts in the supplementary material, with the exceptions listed in the AI-use statement; the result files of Appendices I (3 seeds) and J (4 seeds) are included as stored.