DIAGNOSE BEFORE ADAPTING: INTERVENTION HEADROOM UNDER CLINICAL DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval systems can adapt their evidence memory, decision rule, or model parameters, yet these interventions are often evaluated separately. We study them jointly through *Diagnose-Before-Adapt* (DBA), a prequential intervention audit that measures conditional gains under delayed label release. Recent versus frozen memory measures staleness, cumulative memory measures retention, a support-constrained oracle bounds candidate-restricted decisions, and learned aggregation measures realized gain. Together, these probes characterize the interactions among adaptation components. We apply DBA to 117328 consultations over 51 months and introduce Learned Neighbourhood Aggregation (LNA), a lightweight ranker over prior evidence with frozen text encoders. Fresh-memory headroom grows by more than five balanced-accuracy points per year, while retaining labelled history recovers 20.33–34.98 points. Against recalibrated cumulative memory, base LNA gains 2.83, 1.05, and 3.53 points on the three streams, with over 100$\times$ lower reported GPU occupancy than monthly retraining. Gains are positive in 44 of 45 correlated seed–origin conditions and under label-blind overlap filtering. Public TCM-SD further replicates the aggregation design on a static task. Together, these results identify memory retention and learned evidence use as effective, resource-efficient adaptation components in the studied longitudinal streams.

1 INTRODUCTION

Clinical prediction systems are validated once but deployed into populations, vocabularies, and practices that continue to change. Temporal degradation is well documented (Finlayson et al., 2021; Gama et al., 2014), yet adaptation is usually framed as a parameter update. Retrieval systems expose other levers—refreshing or retaining evidence, recalibrating scores, and learning how to combine retrieved neighbours. Updating everything cannot reveal where actionable headroom remains and may spend GPU compute when labelled history already contains the needed evidence.

We therefore ask: *which intervention has headroom under the deployment timeline?* Existing studies quantify decay or compare update algorithms but often fix memory or aggregation, leaving stale evidence, missing support, and poor use of retrieved support unresolved on the same future cases. We formulate *Diagnose-Before-Adapt* (DBA): estimate the conditional gain and resource increment of changing one component from a matched reference state. This retrospective audit places the accuracy and resource trade-offs of executed interventions on a common deployment timeline.

We study 117328 consultations over 51 months: state S_{t-1} predicts month t, labels are released for scoring, and only afterwards form S_t. DBA measures freshness, retention, support-constrained opportunity, and realized decision gain against explicit reference states under this shared prequential protocol.

This diagnosis motivates *Learned Neighbourhood Aggregation* (LNA; Figure 1). Frozen encoders retrieve prior-month evidence, and a small ranker maps 14 class-wise support, similarity, rank, agreement, and prior features to a six-way prediction. Memory and the scorer update only after current-month labels arrive.

Fresh-memory headroom widens by 5.60 and 5.38 points per year, while cumulative retention recovers 20.33–34.98 points. Against recalibrated cumulative memory, LNA adds 2.83, 1.05, and 3.53

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.

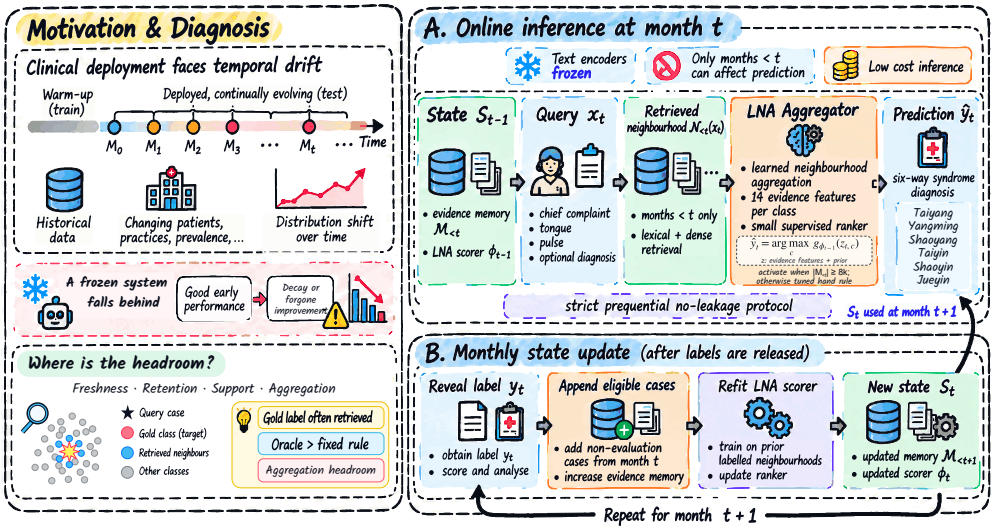


Figure 1: DBA overview. Left: temporal drift motivates separate probes of freshness, retention, retrieval support, and aggregation headroom. Right: at month t, frozen encoders and prior state S_{t-1} use only evidence from months $< t$ to produce the six-class prediction $\hat{y}_t$; after y_t is released, eligible non-evaluation cases update memory and refit the LNA scorer to form S_t for month $t+1$.

points using 0.0020–0.0111 reported GPU-hours, versus 2.28–2.54 for LoRA. The sensitivity grid and overlap controls retain relative gains, and public TCM-SD replicates the design. Exploratory routing further examines the relationship between measured headroom and subsequent action selection.

Our contributions are: (1) a matched prequential audit of memory, decision-rule, and parameter interventions with separate GPU/CPU accounting; (2) a support-constrained oracle ceiling and an interaction analysis that connect retrieval opportunity to conditional intervention gains; and (3) a 51-month empirical study demonstrating the value of a lightweight supervised decision layer over recalibrated cumulative memory, supported by component controls, robustness analyses, and public-task replication.

2 Related Work

Dataset shift in clinical machine learning. Clinical studies document calibration drift (Davis et al., 2017), develop feature-based mitigation (Nestor et al., 2019), and evaluate domain generalisation and unsupervised adaptation (Guo et al., 2022). In the latter study, only unsupervised adaptation uses target-period unlabelled samples. Finlayson et al. (2021) discuss monitoring and mitigation in deployment. We compare memory, decision-rule, and parameter interventions on one delayed-label stream, with matched evaluation and resource accounting.

Concept drift and continual learning. Concept-drift research covers window-based adaptation (Gama et al., 2014; Lu et al., 2019) and adaptive ensembles (Brzezinski & Stefanowski, 2014); our retention controls also include reservoir sampling (Vitter, 1985). Continual learning contributes regularisation (Kirkpatrick et al., 2017), episodic memory (Lopez-Paz & Ranzato, 2017) and exemplar replay (Rebuffi et al., 2017), motivated by catastrophic forgetting (McCloskey & Cohen, 1989). Recent clinical resources expose complementary shift regimes: Clinical-ShiftEval simulates evolving label sets and task definitions (Villena et al., 2026), while MedCL-Bench streams ten biomedical NLP datasets across tasks and measures stability–efficiency trade-offs (Zeng et al., 2026). We study natural temporal shift within one clinical service with a fixed six-class task. The four-year archive is only a few hundred megabytes, allowing us to compare curated and cumulative memory under a stable deployment contract.

Retrieval, stacking, and learned local decisions. After training, predictors can incorporate new information by updating non-parametric memory without retraining their encoders (Lewis et al., 2020; Khandelwal et al., 2020). Learning to combine predictors or local evidence is longstanding: stacked generalisation learns a meta-predictor (Wolpert, 1992), dynamic classifier selection conditions decisions on local competence (Woods et al., 1997), and learning-to-rank optimises ordering from supervised features (Liu, 2009). In TCM decision support, ViTCM-LLM combines multimodal reasoning with retrieval-augmented generation (Luo et al., 2025). Building on tabular ranking over class-wise retrieval statistics, LNA fits its decision layer under a strict prior-month state. Our study connects this learned evidence use to freshness, cumulative memory, calibration, and language-model updates through matched longitudinal cases, paired inference, and resource accounting.

Label shift and calibration. Our calibration layer draws on prior-shift correction (Saerens et al., 2002; Lipton et al., 2018; Alexandari et al., 2020) and post-hoc calibration (Guo et al., 2017). With complete memory, monthly re-estimation changes performance little; this directs the measured opportunity toward learned evidence use.

3 METHOD

3.1 DATA AND TASK

We reuse the longitudinal clinical corpus introduced in ViTCM-LLM (Luo et al., 2025). It contains 117328 de-identified consultations from 2021-04 to 2025-06 with 12 recording physicians. Each record carries a chief complaint, tongue and pulse findings, an optional Western diagnosis, a prescription, and a six-way syndrome label from the *liu-jing* system: 太阳, 阳明, 少阳, 太阴, 少阴, 厥阴. Whereas ViTCM-LLM studies multimodal retrieval-augmented clinical decision support, we use the structured longitudinal records to study temporal adaptation under delayed labels. The label set is complete and mutually exclusive, defining a stable closed-set deployment task. The unit of evaluation is the consultation, and the target is agreement with recorded syndrome labels under temporal deployment.

We use three streams. **HOSP-Stream-Full** is the deployment setting: all physicians, 51 months. **HOSP-Stream-SingleMD** restricts to one physician present in all 51 months (26099 cases), holding the annotator constant. **HOSP-Stream-IndependentMD** is a disjoint physician covering 50 months (79535 cases); the two physician-specific streams share no case. Full contains both, so the trajectories are deployment regimes rather than independent replications. The physician-heterogeneity audit includes the four of 12 physicians meeting outcome-blind thresholds of 1,000 usable cases, 24 observed months, and eight scoreable post-warm-up months. We additionally use **REAL-Meridian-Verified** as an audited static meridian probe and public **TCM-SD** as a separate 148-way task replication of the aggregation design. Appendix B documents every inclusion and exclusion criterion.

3.2 PREQUENTIAL EVALUATION

The first 12 months form a warm-up window; confirmatory hyperparameters are then frozen. Algorithm 1 applies every intervention arm to identical monthly cases and preserves the delayed-label deployment order.

Algorithm 1: Prequential intervention-headroom estimation

1. Initialize matched intervention states from warm-up data.
2. For each month t, freeze every S_{t-1} and predict the same sequestered cases.
3. Release y_t; score paired contrasts and, on probe months, retrieval support.
4. Record GPU/CPU increments; update only with eligible non-evaluation cases from t.
5. Return month-equal headrooms, uncertainty, and the two-resource Pareto profile.

Evaluation cases remain sequestered from every later memory and training pool. Each state records its latest observed month, which the evaluator asserts is earlier than the month being scored. The algorithm returns paired estimates of realized headroom; exploratory routing is evaluated separately in Appendix G.

Definition 1 (Conditional intervention headroom)**.** Let z be a deployed state, I an intervention that changes one designated component, and B_{szt} its balanced accuracy on stream s in month t. Define

$$V_s(z) = |T_s|^{-1} \sum_{t \in T_s} B_{szt}, \qquad H_s(I \mid z) = V_s(I(z)) - V_s(z),$$

with resource increment $\Delta C_s(I \mid z) = (\Delta\mathrm{GPU}, \Delta\mathrm{CPU})$. All other components, evaluation cases, and months are matched. Hence headroom is signed, conditional on z, and path-specific; GPU and CPU are not collapsed into one cost.

Months, not cases, receive equal weight. Balanced accuracy is the mean of six class recalls; plain accuracy instead weights recalls by the target month's natural prior. Abstentions count as errors. Inference resamples months, and pre-specified contrasts use Bonferroni correction with a family size of 63 to address multiplicity. Temporal-dependence sensitivity is reported separately for exploratory LNA+ (Appendix H); base-LNA primary inference uses the stated month-resampling procedure.

Resource accounting. GPU-hours are elapsed exclusive device occupancy and CPU core-hours are process CPU time summed across threads over the evaluation horizon. GPU ratios refer specifically to device occupancy; CPU fitting costs are reported alongside them. Diagnostic arms are executed as part of the evaluation. Method m Pareto-dominates m' when $V_s(m) \geq V_s(m')$, both reported resource increments are no larger, and at least one inequality is strict.

3.3 Diagnose-Before-Adapt framework

Applying Definition 1, we measure freshness $H_s(\text{recent} \mid \text{frozen})$, retention $H_s(\text{cumulative} \mid \text{frozen})$, and realized aggregation $H_s(\text{LNA} \mid \text{BL2b})$, where BL2b is recalibrated cumulative memory. This is a system-level contrast; component ablations isolate dense retrieval and scorer refitting. Every-third-month probes measure retrieval opportunity with fixed memory and retrievers, separately from the full-horizon BL2b comparison.

For query x, let $R_t(x)$ be the union of labels returned by the fixed lexical and dense retrievers, and define the retrospective support oracle $o_t(x, y) = y$ if $y \in R_t(x)$ and $o_t(x, y) = \perp$ otherwise.

Proposition 1 (Support-constrained oracle ceiling)**.** *For any selector h restricted to $R_t(x)$, its month-t balanced accuracy satisfies*

$$B_{st}(h) \leq B_{st}(o) = |\mathcal{C}|^{-1} \sum_{c \in \mathcal{C}} \Pr[c \in R_t(X) \mid Y = c].$$

The inequality is preserved by month averaging and by subtracting the same fixed baseline.

Indeed, a candidate-restricted selector can be correct only when the gold label belongs to $R_t(x)$, while o is correct on every such case. This ceiling is computed after label release and does not bound LNA, which scores all six classes. We report its paired gap from fixed aggregation as support-constrained oracle headroom. A complementary stress test removes gold-supporting neighbours while leaving model fitting and all remaining evidence unchanged.

3.4 Learned Neighbourhood Aggregation

As shown in Figure 1, online inference keeps the text encoders frozen and uses only the prior state S_{t-1}. Within that state, LNA replaces fixed voting and manually tuned score fusion with supervised aggregation over labelled neighbourhoods; the scorer is refitted only after the current month's labels are released. Its default representation has one row for each (query, candidate class) pair and 14 features: lexical support mass and count, best and mean similarity, reciprocal rank, relative similarity and overlap; analogous dense-retrieval quantities; the empirical class prior; and lexical–dense agreement. A small gradient-boosted model ranks the six candidate classes. Formally, let $A_{<t}$ be the fixed-size, class-stratified sample of 3000 labelled cases from months before t. For query q, pre-t neighbourhood $N_{<t}(q)$ and class prior $\pi_{<t}$,

$$\hat{y}_t(q) = \arg\max_{c \in \mathcal{C}} g_{\phi_{<t}}(\psi(q, c, N_{<t}(q)), \pi_{<t}(c)),$$

$$\phi_{<t} = \arg\min_{\phi} \sum_{(x,y) \in A_{<t}} \ell\Big(g_\phi(\psi(x, \cdot, N_{<t}^{(-x)}(x)), \pi_{<t}), y\Big).$$

Each neighbourhood becomes six fixed-dimensional rows: frozen encoders supply similarities, while $\phi_{<t}$ learns the decision rule without encoder back-propagation. The feature map ψ summarizes each retriever separately by class, preserving complementary support and disagreement. Classes without neighbours retain prior features and receive zero support, so LNA scores the full label set. During refitting, $N^{(-x)}$ excludes the training record itself and stratification represents rare labels. Here $< t$ is the deployment cutoff for both the historical training pool and its neighbourhoods, distinct from each record's original month; evaluation-month labels remain unavailable to S_{t-1}. An 84-dimensional representation and a multiclass softmax objective retain the qualitative advantage (Section 4.3).

The scorer is refitted monthly as the eligible retrieval pool accumulates. Held-out warm-up months select a minimum pool of 8000 cases on all three streams; below this threshold, the system uses the tuned hand aggregator. LNA+ is an exploratory long-tail variant selected exclusively on TCM-SD validation: it expands the fitting sample to 12000 neighbourhoods, weights positive class rows, adds a lexical top-1 residual, and applies log-prior correction. The retrievers and 8k activation rule remain unchanged (Appendix D).

Base LNA and the 63-test family are confirmatory; LNA+, routing, component sensitivity, and non-inferiority are exploratory.

4 Experiments

4.1 From temporal drift to fresh-memory headroom

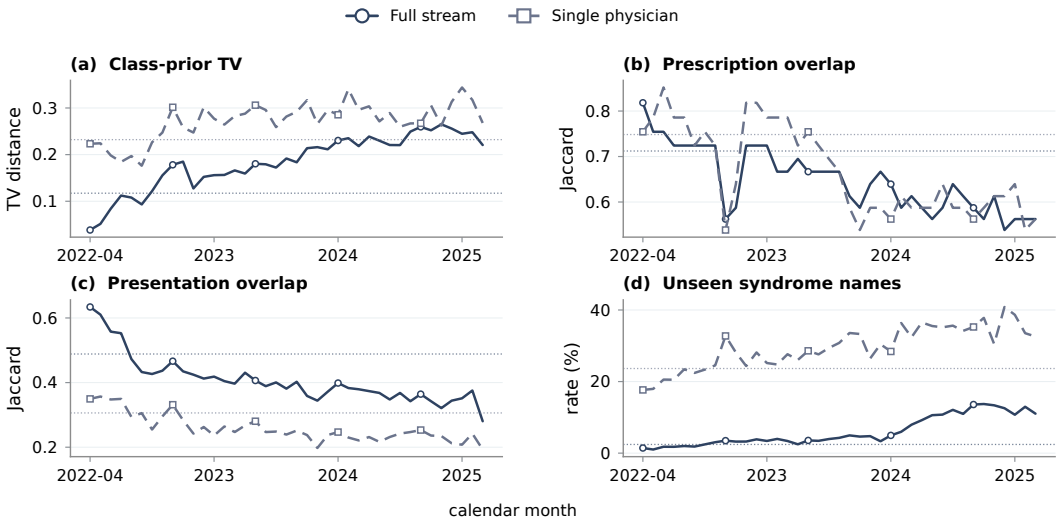


Figure 2: Monthly drift statistics in their original units. Solid circles denote the full stream and dashed squares the single-physician stream; dotted horizontal lines show the corresponding warm-up means.

Figure 2 reports four model-free statistics: total variation of the six-class prior against the warm-up window, Jaccard overlap of the top-50 prescribed herbs, the same for the top-500 retrieval features, and the fraction of cases whose open-vocabulary syndrome name never appeared during warm-up. All four show significant temporal trends; exact slopes are in Appendix Table 4.

Changing physician composition could contribute to pooled drift. The constituent single-physician stream addresses that explanation directly: within one clinician's own case-load, prescription overlap changes at -0.0767 per year (versus -0.0610 pooled) and unseen syndrome names at +0.0536 per year (versus +0.0398 pooled). Comparable proxy drift is therefore present even with the recording physician fixed.

We next turn from proxy drift to its predictive consequence using a size-controlled memory comparison.

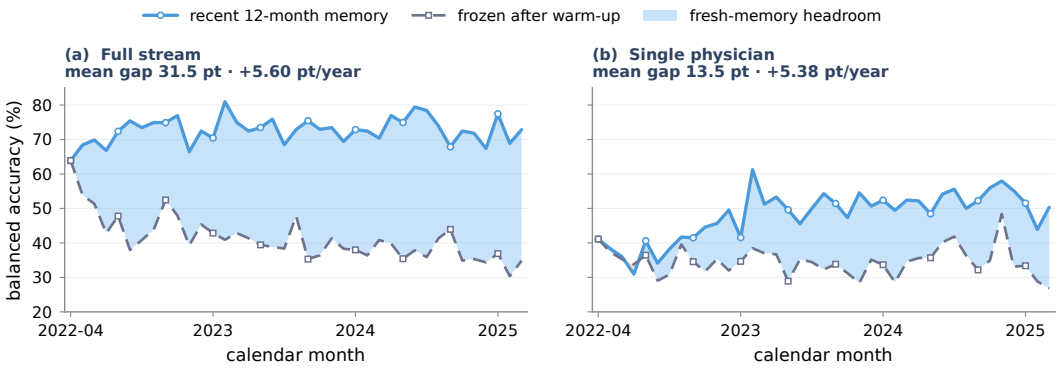


Figure 3: Fresh-memory headroom over time. The shaded ribbon is the paired gap between a size-matched recent 12-month memory and the memory frozen after warm-up; orange marks the rare months where the ordering reverses. Insets report the mean gap and its annual trend.

We compare a system frozen after warm-up against one whose memory is rebuilt from the twelve months preceding each evaluation month. The comparison is *size-matched*: monthly case volume grows over the period, so an unmatched recent window would hold several times more evidence and confound recency with memory size. Both arms hold exactly the same number of cases but draw them from different time periods. As a check, the arms agree exactly at the first evaluation month, where the memories coincide. Exact means and trends are reported in Appendix Table 5.

On HOSP-Stream-Full the frozen system falls from 47.29 to 36.74, a trend of -4.90 points per year ($p < 10^{-4}$, $\rho_t = -0.74$), while the fresh control is flat (+0.69 per year, $p = 0.336$). On HOSP-Stream-SingleMD the frozen system is flat (-0.30 per year, $p = 0.723$) while the fresh control *rises* at +5.08 per year.

Both streams exhibit a widening predictive-performance gap, although its level differs: +31.47 points on Full and +13.53 on SingleMD. The estimated annual slopes are +5.60 and +5.38 points. In the pooled stream the gap appears as absolute decay; in its constituent single-physician stream it appears as forgone improvement. The paired, size-matched gap is therefore the shared estimand, without requiring the two trajectories to have the same cause or magnitude.

4.2 DBA INTERVENTION LEDGER

The dominant non-parameter intervention is memory retention. Appending eligible cases raises balanced accuracy by 34.98, 20.33, and 33.03 points on Full, SingleMD, and IndependentMD, respectively (all $p < 10^{-4}$), and reverses the Full-stream trend from 73.29 to 78.96. Retaining the archive also beats the tested bounded alternatives: a six-month window loses 3.32 and 4.57 points on Full and IndependentMD, while a 20000-case reservoir loses 7.46 and 5.68. Older cases therefore remain useful in these streams, where retaining the eligible archive is practical.

Calibration and language-model fusion contribute little once memory is complete. Monthly recalibration changes accuracy by -0.17 points ($p = 0.4316$), with stream effects from -0.17 to +1.33; log-linear fusion changes accuracy by -0.16, +0.68, and +0.33 points, and warm-up fitting selects zero LM weight. LoRA adds 2.32, 0.51, and 4.85 points over its matched no-gradient system, but its detector fires in 39/39, 35/39, and 38/38 months and occupies 2.54, 2.28, and 2.52 GPU-hours. We next ask how much opportunity remains inside the retrieved evidence. At $k = 32$, class-balanced lexical support is 93.41%, 75.67%, and 91.07%; the lexical–dense support oracle reaches 95.96%, 79.73%, and 93.67% balanced accuracy while the fixed rule reaches 74.79%, 53.58%, and 71.53%. Proposition 1 therefore yields 21.17–26.14 points of support-constrained headroom. Separately, base LNA gains 1.05–3.53 points over BL2b across the full evaluation horizon. Oracle headroom characterizes candidate-restricted opportunity on probe months, whereas the BL2b comparison measures realized system gain across evaluation months.

Proposition 2 (Path dependence under interaction)**.** *For performance $V_{ma} = V(M_m, A_a)$ under two memory states and two decision rules, let $I = V_{11} - V_{10} - V_{01} + V_{00}$. Both intervention paths*

Table 1: Confirmatory comparison over evaluation months. The blue row marks our base LNA; bold is the best overall accuracy and underline is the best result without language-model parameter updates. Resource columns report the range across the three streams; GPU and CPU remain separate quantities.

Method	Full	SingleMD	IndependentMD	GPU h	CPU core-h
Memory and calibration					
BL1 Frozen	41.21	34.55	39.19	0	0.001
BL1b Frozen + recalibration	42.00	32.09	39.18	0	0.115–0.170
BL2 Cumulative	76.20	54.87	72.22	0	0.005–0.042
BL2b Cumulative + recalibration	76.03	56.20	72.86	0	0.121–0.210
BL3 Sliding window	72.88	52.98	67.66	0	0.002–0.010
BL4 Reservoir replay	68.73	54.87	66.55	0	0.005–0.014
CDA-L1 (memory)	75.82	54.87	72.62	0	0.005–0.038
CDA-L2 (calibration)	76.03	56.20	72.86	0	0.122–0.209
CDA-L1L2	76.52	56.20	72.48	0	0.123–0.203
Frozen language-model inference and parameter adaptation					
Frozen LM + frozen memory	40.53	34.65	38.64	0.115–0.121	0.002–0.003
CDA-L1L2 + frozen LM	76.36	56.88	72.81	0.114–0.122	0.008–0.041
CDA-L3 (LoRA adaptation)	78.68	57.40	**77.66**	2.284–2.543	0.010–0.046
BL5 Monthly full-history retrain	77.37	**58.45**	76.01	2.564–2.641	0.011–0.046
Learned neighbourhood aggregation					
LNA (lexical only)	77.84	56.46	74.74	0	0.033–0.124
LNA (scorer fitted once)	77.92	56.95	73.02	0.002–0.012	0.014–0.065
LNA (base, ours)	**78.85**	57.25	76.39	0.002–0.011	0.053–0.207

Table 2: Selected pre-specified contrasts on monthly balanced accuracy, paired by month and Bonferroni-corrected with family size 63; ** marks survival at 0.05 under the stated month-resampling procedure.

Contrast	HOSP-Stream-Full	HOSP-Stream-SingleMD	HOSP-Stream-IndependentMD
memory: cumulative vs frozen	+34.98**	+20.33**	+33.03**
calibration alone	+0.78	-2.45**	-0.01
calibration on top of memory	-0.17	+1.33	+0.64
window vs cumulative	-3.32**	-1.89**	-4.57**
reservoir vs cumulative	-7.46**	+0.00	-5.68**
CDA-L1 vs cumulative	-0.38	+0.00	+0.40
CDA-L1L2 vs recalibrated cumulative	+0.50	+0.00	-0.38
Frozen-LM fusion vs CDA-L1L2	-0.16	+0.68	+0.33
L3 vs CDA-L1L2 + frozen LM	+2.32**	+0.51	+4.85**
L3 vs monthly full-history retrain	+1.32**	-1.06	+1.65**
LNA vs recalibrated cumulative	+2.83**	+1.05	+3.53**
LNA vs cumulative	+2.66**	+2.37**	+4.16**
LNA vs CDA-L1L2	+2.33**	+1.05	+3.91**
LNA vs LoRA adaptation	+0.17	-0.15	-1.27
LNA vs monthly full-history retrain	+1.49	-1.20	+0.38
LNA: dense retriever ablation	+1.01	+0.79	+1.64**
LNA: monthly refit ablation	+0.93	+0.30	+3.37**

telescope to $V_{11} - V_{00}$, but each component's conditional effect changes by I. Thus component attribution is order-invariant if and only if $I = 0$.

This follows by expanding $(V_{11} - V_{01}) - (V_{10} - V_{00}) = (V_{11} - V_{10}) - (V_{01} - V_{00}) = I$. Figure 4 reports one sequential path. The exploratory memory–LNA+ factorial provides an empirical instance with interactions of -9.86, -8.44, and -7.81 points (Appendix F), quantifying how the memory state changes the gain available to learned aggregation.

Appendix Fig. 14 reports the controlled support test. Pooling probe cases within class gives natural union support of 95.96%, 84.60%, and 93.64%; the oracle values above instead average class-balanced accuracy equally over months. Removing all gold-supporting neighbours without refitting reduces LNA+ to 4.80%, 13.01%, and 8.11%. Enlarging neighbourhoods raises support without

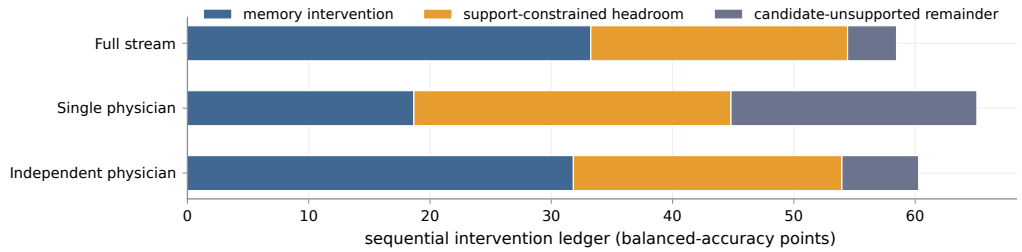


Figure 4: Sequential intervention ledger on every-third-month probes. Segments report the memory intervention, support-constrained oracle headroom, and the candidate-unsupported remainder. They telescope within each stream but depend on intervention order.

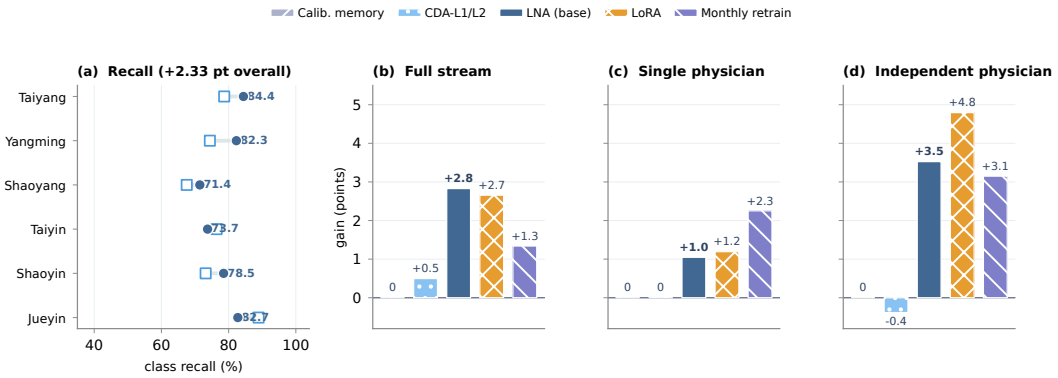


Figure 5: Confirmatory base-LNA performance. (a) Paired six-meridian recall on Full, averaged equally over evaluation months. (b–d) Gain over calibrated cumulative memory for five matched policies on each stream. The primary comparison uses the pre-specified base LNA; absolute accuracy and separate GPU/CPU costs are in Table 1, and Figure 6 reports the complete frontier.

consistently raising accuracy (Appendix G). Together, these interventions show that retrieved support matters under the specified perturbation, but additional support alone does not determine utility.

An exploratory leave-one-stream-object-out test asks whether these diagnostics select a future action (Appendix G). Router accuracy is 79.39%, 58.63%, and 76.81%: it matches the best always-on policy on Full and SingleMD, but remains 1.83 points below always-LoRA on IndependentMD and matches hindsight there on 1/12 targets. This stream-object split retains the case and physician overlap between Full and its constituent streams. The results distinguish within-stream intervention gains from the stream-specific behaviour of the routing rule.

4.3 LNA PERFORMANCE, EFFICIENCY, AND ROBUSTNESS

Base LNA has the highest point estimate among the Table 1 methods without language-model parameter updates. Relative to the named BL2b reference, it gains 2.83, 1.05, and 3.53 points; the Full and IndependentMD contrasts survive the reported correction, whereas SingleMD does not. On SingleMD, frozen-LM fusion reaches 56.88, compared with LNA's 57.25, a displayed difference of about 0.37. Against plain accumulation, gains are 2.66, 2.37, and 4.16; Table 1 gives absolute accuracy and both resource budgets.

Across five evaluation draws and 6-, 12-, and 18-month warm-up boundaries, mean gains over recalibrated cumulative memory are 2.61, 1.06, and 3.52 points, positive in 15/15 Full, 14/15 SingleMD, and 15/15 IndependentMD conditions in the correlated sensitivity grid (Appendix Fig. 10). Label-blind exact/high-similarity filtering affects 46.08%–61.21% of queries, with positive relative gains on Full (4.09/4.36), SingleMD (2.16/2.11), and IndependentMD (4.84/4.78). Absolute LNA accuracy under high-similarity filtering is 57.72, 43.29, and 55.92, lower than the unfiltered scores in

Table 3: Base-LNA sensitivity to identical label-blind overlap removal before retrieval.

Stream	Exclusion	Queries matched	Baseline	LNA	LNA gain [95% CI]
HOSP-Stream-Full	Normalized exact	59.76	53.84	**57.92**	**+4.09 [+2.83, +5.35]**
HOSP-Stream-Full	High similarity	61.21	53.36	**57.72**	**+4.36 [+3.09, +5.63]**
HOSP-Stream-SingleMD	Normalized exact	46.08	41.41	**43.57**	**+2.16 [+1.13, +3.19]**
HOSP-Stream-SingleMD	High similarity	48.43	41.18	**43.29**	**+2.11 [+1.00, +3.24]**
HOSP-Stream-IndependentMD	Normalized exact	54.39	51.30	**56.14**	**+4.84 [+3.56, +6.15]**
HOSP-Stream-IndependentMD	High similarity	55.54	51.14	**55.92**	**+4.78 [+3.50, +6.09]**

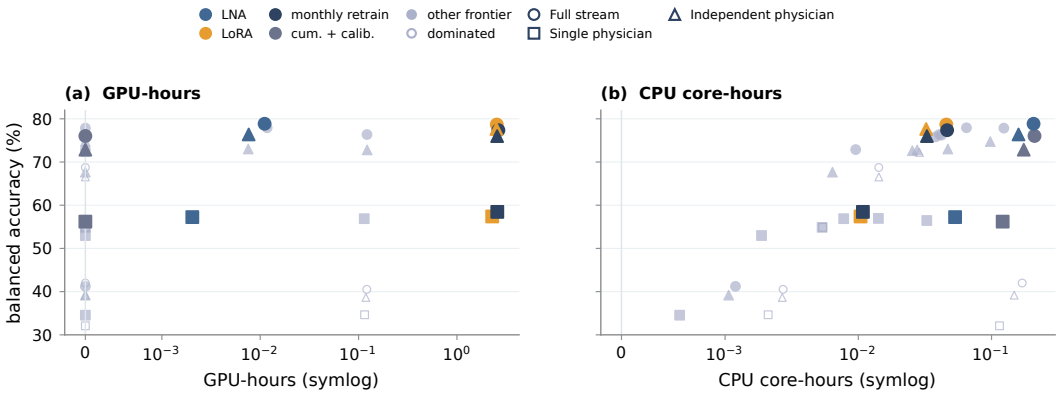


Figure 6: Accuracy–resource frontier over all evaluated arms. GPU occupancy and CPU fitting cost remain separate; filled points are Pareto-optimal and hollow points are dominated.

Table 1. Table 3 thus reports both the effect of removing similar evidence and the retained aggregation gain.

Relative to parameter-updated references, LNA differs from LoRA by +0.17, -0.15, and -1.27 points, and from monthly retraining by +1.49, -1.20, and +0.38; none of the LoRA contrasts survives the 63-test correction (Appendix Table 14). Monthly retraining uses 2.56–2.64 GPU-hours versus LNA's 0.0020–0.0111, placing base LNA on every measured GPU/CPU Pareto frontier (Figure 6).

Four representation–objective variants outperform the fixed rule; a rolling prior explains only 6.4%–38.7% of the learned gain (Appendix C).

Figure 6 separates two operational choices: allocating GPU time to update language-model parameters, and allocating CPU time to fit a decision rule over frozen representations. LNA occupies the low-GPU region while using CPU fitting to improve on calibrated cumulative memory. The frontier therefore makes the accuracy–resource trade-off explicit for each stream. Read alongside Table 3 and the representation ablations, this profile links the measured gain to learning how to combine available evidence. Memory retention and learned aggregation contribute at different stages of the same pipeline, making both relevant when comparing adaptation options under a given resource budget.

5 Conclusion

DBA brings memory, decision-rule, and parameter interventions onto a shared delayed-label timeline. Across 51 months, cumulative memory recovers 20.33–34.98 points over frozen memory; base LNA adds 2.83, 1.05, and 3.53 over recalibrated cumulative memory, with low reported GPU occupancy. Relative gains persist across the seed–origin grid and label-blind overlap controls, while public TCM-SD replicates the aggregation design on a distinct task. These findings establish the value of retaining labelled history and learning how to use it in the studied streams. The resulting intervention ledger makes the roles of evidence freshness, retained support, and learned decisions explicit, enabling resource-aware comparison of temporal adaptation components.

Artificial Intelligence (AI) Use Statement

An LLM assisted with manuscript drafting, substantive revision, copy-editing, and research support, including literature discovery, brainstorming and stress-testing the intervention-headroom framing, pre-submission review of claims and evidence, and LaTeX revision and compilation checks. Such assistance did not generate new experimental observations. Most relevantly, discussion with an LLM helped us articulate the distinction between measuring conditional intervention headroom and selecting a prospective adaptation policy, which organises the present framing. All conceptual contributions and experimental design decisions were made by the authors. All reported results are drawn from the experimental artefacts described in Appendix J. The authors are responsible for verifying the references, experimental claims, and final wording.

Reproducibility Statement

The anonymous bundle contains implementation code, frozen configurations, regression tests, aggregate result artefacts, and all table and figure builders. Appendix J records the executable configuration and determinism checks. The private clinical corpus and model weights are not redistributed; end-to-end reruns require the documented corpus, encoder, and language-model resources. All reported tables and figures can be regenerated from the released aggregate artefacts.

References

Amr M. Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232. PMLR, 2020. URL https://proceedings.mlr.press/v119/alexandari20a.html.

Dariusz Brzezinski and Jerzy Stefanowski. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 25 (1):81–94, 2014. doi: 10.1109/TNNLS.2013.2251352.

Sharon E. Davis, Thomas A. Lasko, Guanhua Chen, Edward D. Siew, and Michael E. Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017. doi: 10.1093/jamia/ocx030.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021. doi: 10.1056/NEJMc2104626.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014. doi: 10.1145/2523813.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12(1):2726, 2022. doi: 10.1038/s41598-022-06484-1.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

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, 2017. doi: 10.1073/pnas.1611835114.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. doi: 10.1561/1500000016.

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/7225-gradient-episodic-memory-for-continual-learning.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi: 10.1109/TKDE.2018.2876857.

Lihui Luo, Joongwon Chae, Yang Liu, Igor Pantic, Vladan Devedzic, Zhumei Sun, Zelin Zeng, Sanatbek Matlatipov, Xiaoming Yin, and Peiwu Qin. ViTCM-LLM: A multimodal RAG framework for advanced TCM clinical decision support. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3894–3899. IEEE, 2025. doi: 10.1109/BIBM66473.2025.11357113.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower (ed.), *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Elsevier, 1989. doi: 10.1016/S0079-7421(08)60536-8.

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*, volume 106 of *Proceedings of Machine Learning Research*, pp. 381–405. PMLR, 2019. URL https://proceedings.mlr.press/v106/nestor19a.html.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542. IEEE, 2017. doi: 10.1109/CVPR.2017.587.

Mucheng Ren, Heyan Huang, Yuxiang Zhou, Qianwen Cao, Yuan Bu, and Yang Gao. TCM-SD: A benchmark for probing syndrome differentiation via natural language processing. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pp. 908–920. Chinese Information Processing Society of China, 2022. URL https://aclanthology.org/2022.ccl-1.80/.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

Fabián Villena, Felipe Bravo-Marquez, and Jocelyn Dunstan. Clinical-ShiftEval: A framework for simulating and evaluating model adaptation in dynamic clinical NLP tasks. *BMC Medical Informatics and Decision Making*, 26(1):240, 2026. doi: 10.1186/s12911-026-03538-6.

Jeffrey S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985. doi: 10.1145/3147.3165.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Kevin Woods, W. Philip Kegelmeyer, and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (4):405–410, 1997. doi: 10.1109/34.588027.

Min Zeng, Shuang Zhou, Zaifu Zhan, and Rui Zhang. MedCL-Bench: Benchmarking stability–efficiency trade-offs and scaling in biomedical continual learning. *arXiv preprint arXiv:2603.16738*, 2026. doi: 10.48550/arXiv.2603.16738.

A ADDITIONAL TEMPORAL DIAGNOSTICS

Drift slopes and freshness-control estimates. Table 4 gives the exact annual trend estimates underlying Figure 2.

Table 4: Model-free drift diagnostics relative to warm-up. Label-prior and syndrome-name statistics are computed from released annotations. Both streams have reported temporal trends on every statistic ($p < 10^{-4}$); within-physician trends argue against changing physician composition as the sole explanation.

Statistic	Stream	Year 1	Year 4	Per year	ρ_t
Label prior (total variation)	Full	0.117	0.242	+0.0573	+0.94
	Single-MD	0.232	0.289	+0.0274	+0.61
Prescription top-50 (Jaccard)	Full	0.712	0.582	-0.0610	-0.82
	Single-MD	0.748	0.592	-0.0767	-0.71
Presentation top-500 (Jaccard)	Full	0.488	0.348	-0.0654	-0.93
	Single-MD	0.306	0.229	-0.0358	-0.81
Unseen syndrome names (rate)	Full	0.024	0.118	+0.0398	+0.95
	Single-MD	0.236	0.355	+0.0536	+0.86

Table 5: Frozen and size-matched recent-memory controls. Both arms reuse the exact BL1 warm-up calibration from Table 1; the contrast changes only the temporally refreshed, size-matched memory state.

Stream	Memory	Mean	Year 1	Year 4	Per year	p
HOSP-Stream-Full	frozen	41.21	47.29	36.74	-4.90	0.0000
	fresh, size-matched	72.69	71.33	73.53	+0.69	0.3360
	fresh, 1-NN	71.55	69.63	72.47	+0.67	0.3390
staleness gap		+31.47	+24.04	+36.79	+5.60	0.0000
HOSP-Stream-SingleMD	frozen	34.55	34.75	35.61	-0.30	0.7230
	fresh, size-matched	48.08	40.20	52.29	+5.08	0.0000
	fresh, 1-NN	45.81	38.79	49.79	+4.79	0.0000
staleness gap		+13.53	+5.45	+16.68	+5.38	0.0000

Table 6: External-corpus trajectories and retrospective early-month re-evaluation. The final state improves predictions on held-out cases from early months; this is a retention diagnostic, not a prospective estimate for those months.

Stream	Frozen	Adapted	Per year	Early-month re-eval.
HOSP-Stream-Full	19.92	23.46	+6.44	+18.22
HOSP-Stream-SingleMD	25.49	29.52	+5.81	+20.71
HOSP-Stream-IndependentMD	22.15	27.64	+8.84	+20.40

Retrospective early-month and external re-evaluation. Re-scoring held-out cases from the earliest evaluation months with the final state improves balanced accuracy by +18.22, +20.71 and +20.40 points, with every sampled month positive. The comparison quantifies retention and the value of later evidence for earlier presentations; it is separate from the prequential deployment estimate.

B BENCHMARK AUDIT

Table 7 lists five used evaluation settings and five excluded entries. Exclusions are based on label validity, overlap with the clinical stream, or mismatch with the target temporal-adaptation task.

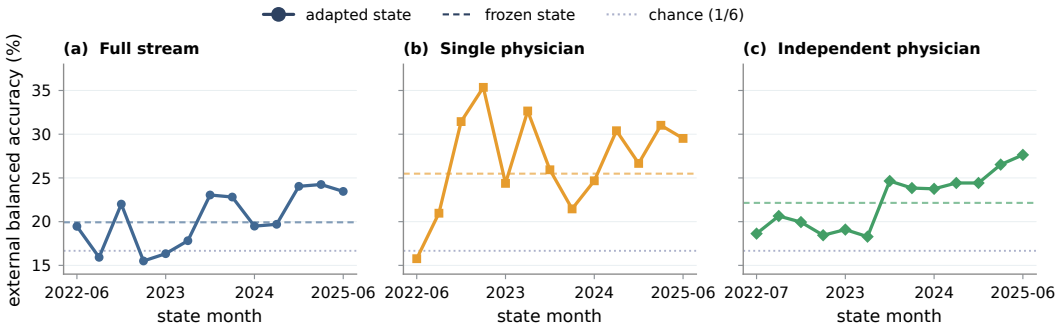


Figure 7: Balanced accuracy on the fixed REAL-Meridian-Verified set as each internal stream state advances. Dashed horizontal lines are the corresponding frozen states; the dotted line marks chance performance.

Table 7: Evaluation settings and excluded benchmark entries. The final row groups two multiple-choice resources with the same task mismatch.

Benchmark	Kind	n	Disposition
HOSP-Stream-Full	monthly stream	117,328	used, 39 evaluation months
HOSP-Stream-SingleMD	monthly stream	26,099	used, 39 evaluation months
HOSP-Stream-IndependentMD	monthly stream	79,535	used, 38 evaluation months
REAL-Meridian (tcm_fd)	static, external	212	used (confidence 1.0)
TCM-SD (original)	static, 148-way	54,152	used, official split
REAL-Meridian (tcm_sd)	static	400	excluded: keyword-generated labels
REAL-Meridian (myd)	static	76	excluded: compound labels collapsed
TCM-SD-derived meridian subset	static	1,500	excluded: no meridian label
HOSP-OOD	static	1,000	excluded: subset of the stream
MTCMB / ShizhenGPT	multiple choice	6,488	excluded: different task format

Table 8: Exploratory component factorial. Balanced accuracy is measured on the same subsampled months for two representations and two learning objectives, beside two hand-designed aggregators. Learning beats the tuned rule in every cell; no representation or objective wins consistently across streams. These post-hoc tests are descriptive and are not part of the confirmatory family in Table 2.

Aggregator	HOSP-Stream-Full	HOSP-Stream-SingleMD	HOSP-Stream-IndependentMD
Tuned hand aggregation	74.78	58.47	73.19
Hand aggregation + adaptive prior	75.03	59.79	73.59
Pairwise learner, relational features	**78.76**	61.90	77.69
Pairwise learner, class-level features	77.47	**64.85**	78.42
Softmax learner, relational features	77.47	60.98	**79.14**
Softmax learner, class-level features	77.05	62.98	77.61

The bundled REAL-MERIDIAN set contains 688 cases from three sources, of which we keep 212. The 400 derived from TCM-SD carry labels generated by keyword rules of the form "*qi deficiency* → 太阴" at confidence 0.7–0.8, and that subset was constructed to be exactly balanced at 80 cases per class. A retrieval system that has seen the triggering phrase paired with the same rule scores well there without establishing clinical generalisation, so accuracy would estimate agreement with the label-generation script. A further 76 cases collapse compound syndromes onto a single meridian, making the target unidentifiable from the input. Separately, the bundled 1500-record TCM-SD-derived subset has an empty meridian field and cannot support the six-way meridian probe; the original 54,152-record TCM-SD release remains usable for its distinct 148-way syndrome task in Appendix D. Finally, a bundled "out-of-distribution" hospital set of 1000 cases is not external — every one of its case identifiers occurs inside our stream.

Two multiple-choice benchmarks were excluded as task-incompatible: performance there reflects general medical knowledge rather than adaptation to a specific practice. In place of a static second transfer probe we use HOSP-Stream-IndependentMD, an independent physician-specific 50-month adaptation trajectory with real labels, timestamps, and two orders of magnitude more data. Because it updates on that physician's own prior months, it is a replication stream rather than cross-physician generalisation.

C COMPONENT AND REPRESENTATION SENSITIVITY

Every supervised objective–representation cell in Table 8 exceeds the tuned hand rule by approximately 2.27–6.38 points; pairwise versus softmax objectives change order across streams. In the separate representation comparison of Table 9, displayed relational-minus-class-level differences are about +0.22, -0.04, and +1.27 points. Taken together, the separate factorial and representation studies show gains across the tested representations and objectives.

The learner combines evidence with a useful prior rather than acting as prior correction alone. A seven-month rolling prior accounts for 6.4%, 38.7%, and 8.8% of the learned gain, while leave-one-group-out ablation identifies the prior as the most consistently useful input: removing it costs 2.25 and 3.83 points on the two large streams, while dense retrieval is material on HOSP-Stream-IndependentMD (1.89 points). These facts are compatible: the learner uses the prior *jointly* with evidence, but merely substituting a newer prior into the old aggregation rule does not reproduce what it learns.

D PUBLIC BENCHMARK EVALUATION

We evaluate an independently published benchmark with no exact normalized-text overlap with the hospital stream. TCM-SD (Ren et al., 2022) contains 54,152 real clinical records and uses the authors' official 43,180/5,486/5,486 split for 148-way syndrome classification. The base LNA already reaches 82.72% accuracy and 56.25% macro-F1. Selecting the long-tail recipe on validation only raises the untouched test result to 83.19% and 58.50%. This exceeds the published RoBERTa accuracy (82.26%) and ZY-BERT macro-F1 (51.01%), while lexical 1-NN remains a strong 80.82%

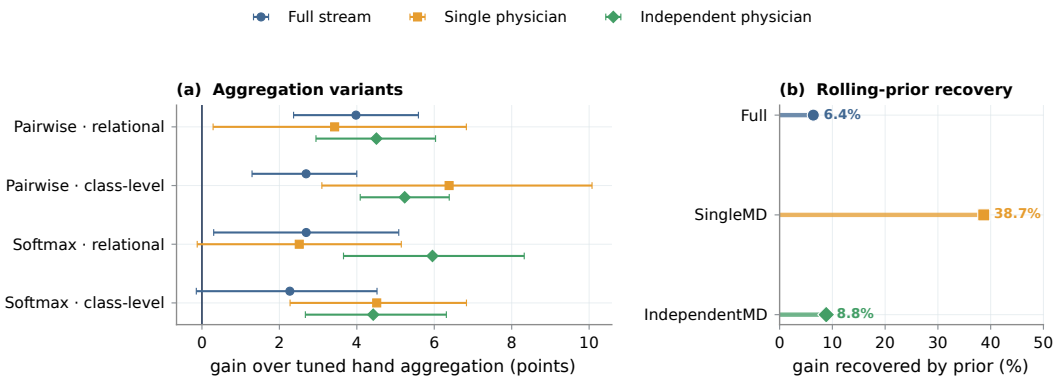


Figure 8: Component sensitivity. Left: learned aggregators relative to tuned hand aggregation, with 95% paired-bootstrap intervals. Right: fraction of the learned gain reproduced by a rolling class prior. Marker shape identifies the stream.

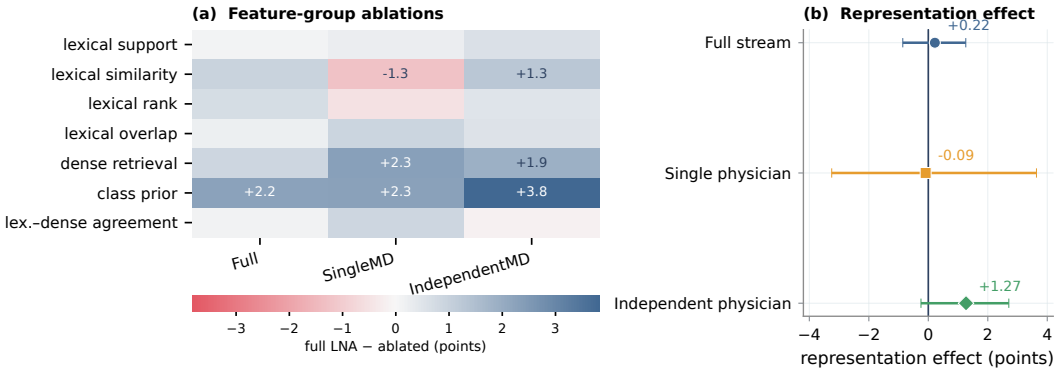


Figure 9: LNA ablations. Left: full-minus-ablated balanced accuracy for seven feature groups; only effects of at least one point are printed. Right: relational-minus-class-level effects with 95% paired-bootstrap intervals.

accuracy baseline. This is an external task replication: LNA is trained and evaluated on the official TCM-SD split, testing the aggregation design rather than transfer of a hospital-trained model.

Table 11 gives the complete listed comparison. TCM-SD matches the paper's retrieval-supported classification setting while changing the corpus and label space.

E Robustness, Alternatives, and Deployment

Evaluation draws and rolling origins. Figure 10 and Table 13 report the $3 \times 5 \times 3$ sensitivity grid. Gains over recalibrated cumulative memory are positive in 44 of 45 trajectories. Draws and

Table 9: Exploratory representation robustness over the sampled evaluation horizon. The uncorrected paired tests find no significant difference between relational and class-level inputs; these tests diagnose mechanism rather than support the confirmatory claim.

	HOSP-Stream-Full		HOSP-Stream-SingleMD		HOSP-Stream-IndependentMD	
Representation	Bal. acc.	p vs. other	Bal. acc.	p vs. other	Bal. acc.	p vs. other
Relational (14-D per class)	**77.74**	0.686	56.23	0.921	**76.83**	0.101
Class-level (84-D per query)	77.52	–	**56.27**	–	75.56	–

Table 10: Exploratory leave-one-group-out ablation of LNA. Entries are full minus ablated balanced accuracy in points; positive values mean the feature group helps.

Dropped feature group (Δ = full − ablated)	HOSP-Stream-Full	HOSP-Stream-SingleMD	HOSP-Stream-IndependentMD
Lexical support	+0.12	+0.27	+0.60
Lexical similarity	+0.97	-1.28	+1.28
Lexical rank	+0.70	-0.54	+0.50
Lexical overlap	+0.23	+0.90	+0.55
Dense retrieval	+0.89	+2.31	+1.89
Class prior	+2.25	+2.30	+3.83
Lexical–dense agreement	+0.13	+0.88	-0.17

Table 11: TCM-SD test results. Published rows are copied from the original papers, not re-run; all LNA variants use the official split.

Benchmark	Method	Accuracy	Macro-F1
TCM-SD	RoBERTa (published)	82.26	47.55
TCM-SD	ZY-BERT (published)	82.19	51.01
TCM-SD	Lexical 1-NN	80.82	54.83
TCM-SD	Dense 1-NN	72.64	43.63
TCM-SD	LNA (base)	82.72	56.25
TCM-SD	**LNA+**	**83.19**	**58.50**

Table 12: Exploratory transport of the LNA+ recipe frozen on TCM-SD validation to the hospital streams. Deltas are balanced-accuracy points. The 12 pairwise comparisons form a separate Bonferroni family and do not alter the confirmatory family.

Stream	Bal. acc.	Δ base	Δ calib.	Δ LoRA	Δ retrain	GPU h / CPU h
HOSP-Stream-Full	**79.57**	+0.72	+3.55	+0.89	+2.21	0.014 / 0.635
HOSP-Stream-SingleMD	**58.39**	+1.14	+2.19	+1.00	-0.06	0.003 / 0.128
HOSP-Stream-IndependentMD	**75.68**	-0.70	+2.82	-1.97	-0.33	0.010 / 0.473

origins reuse overlapping records and months, and Full contains the physician-specific streams. Accordingly, the grid summarizes sensitivity to evaluation draws and boundaries within the same longitudinal archive.

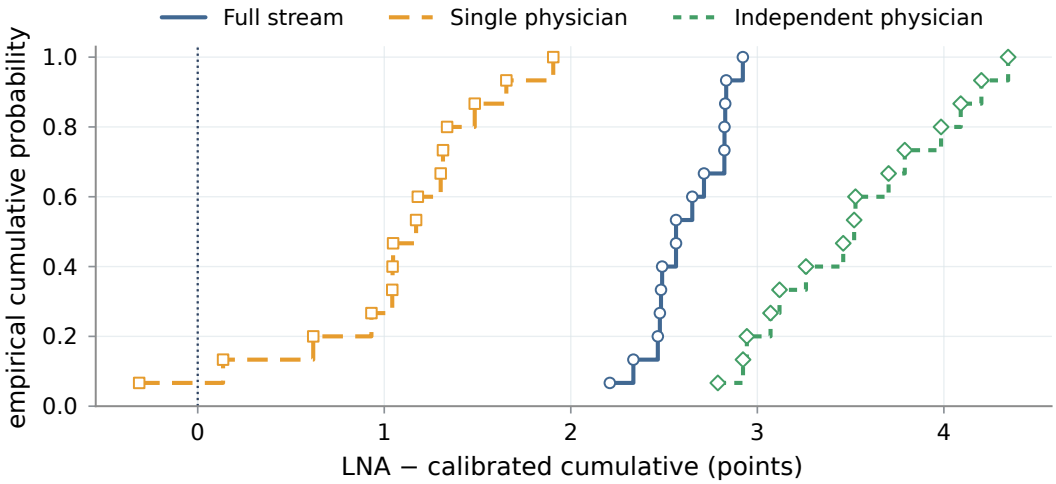


Figure 10: Empirical cumulative distributions of LNA gains over calibrated cumulative memory across 15 complete seed–origin trajectories per stream. Each monthly evaluation set is re-drawn and withheld from all future memories; the vertical line marks zero gain.

Table 13: Robustness over 15 evaluation-seed/origin conditions per stream. SD and range are across complete online trajectories, not case-level resamples.

Stream	Mean Δ	SD	Range	Positive	Rank 1
HOSP-Stream-Full	**+2.61**	0.21	[+2.21, +2.92]	**15/15**	**15/15**
HOSP-Stream-SingleMD	**+1.06**	0.56	[-0.31, +1.91]	**14/15**	**14/15**
HOSP-Stream-IndependentMD	**+3.52**	0.50	[+2.79, +4.34]	**15/15**	**15/15**

Text-overlap sensitivity details. Table 3 reports the paired effects in the main text. The exclusion is query-specific and label-blind: it removes only strictly earlier normalized-exact or high-similarity character-trigram matches, identically before lexical and dense retrieval, while model fitting and the prequential timeline remain unchanged. The control isolates retrieval-time text overlap, with both absolute accuracy and paired gains reported in the table. The evaluation scope is case-sequestered temporal prediction and retrieval-time filtering; patient-disjoint transfer and clinical outcomes are outside this design.

Exploratory non-inferiority sensitivity. Table 14 reports moving-block lower bounds at 1-, 2-, and 3-point margins. The 2-point convention is evaluated across six Bonferroni-corrected comparisons; these post-hoc results describe margin sensitivity and do not alter the confirmatory family.

Physician-level policy estimates. Across four mutually exclusive physician streams, the reported policy contrast has a random-effects mean of 1.73 points. The Hartung–Knapp 95% interval is $[-0.71, +4.17]$ and the prediction interval is $[-5.16, +8.62]$. Two streams remain below the ranker's activation threshold. We therefore distinguish the full-policy estimand from the active-LNA estimand, which uses only months with a fitted ranker (Appendix I).

Alternative supervised aggregators and tail risk. On identical prior-month neighbourhood matrices, histogram-gradient boosting has the highest mean score on all three streams; logistic regres-

Table 14: Post-hoc non-inferiority analysis using a three-month circular moving-block bootstrap. Δ and the one-sided lower bound are for LNA minus the reference, in points. †The 2-point decisions are corrected across six comparisons.

Stream	Reference	Δ	Lower 95%	NI@1pt	NI@2pt†	NI@3pt
HOSP-Stream-Full	LoRA	+0.17	-0.85	Yes	Yes	Yes
HOSP-Stream-Full	Monthly retrain	+1.49	+0.72	Yes	Yes	Yes
HOSP-Stream-SingleMD	LoRA	-0.15	-1.50	No	No	Yes
HOSP-Stream-SingleMD	Monthly retrain	-1.20	-3.01	No	No	No
HOSP-Stream-IndependentMD	LoRA	-1.27	-2.14	No	No	Yes
HOSP-Stream-IndependentMD	Monthly retrain	+0.38	-0.49	Yes	Yes	Yes

sion, a two-layer MLP, and DeepSets provide controlled alternatives on the same evidence. The ordering supports the selected tabular ranker rather than attributing the gain to supervised aggregation alone. Tail metrics separate where LNA+ helps from where more evidence is required: it improves worst-class recall on every stream, while low-coverage novel terminology remains assigned to the retrieval branch of the diagnostic.

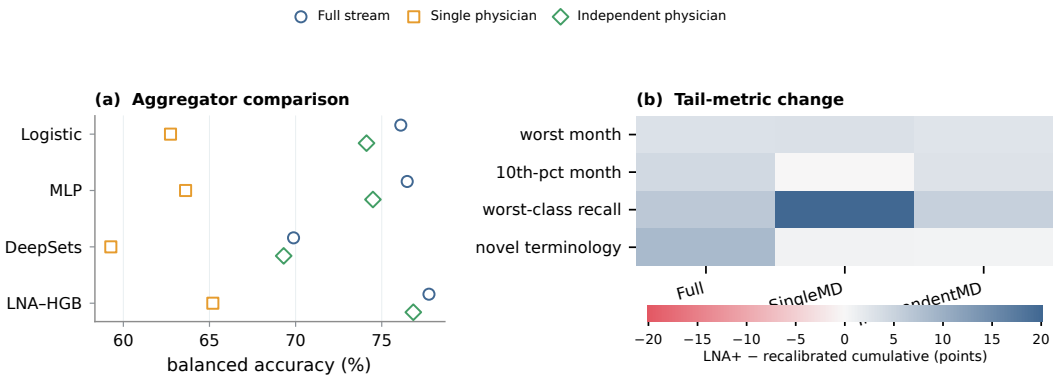


Figure 11: Alternative supervised aggregators on identical neighbourhood matrices (left) and LNA+ tail-metric changes relative to recalibrated cumulative memory (right). Marker shape identifies the stream; the heatmap is centred at zero.

Deployment profile. At 107,136 memory cases, warm in-process inference has 23.8-ms median and 25.0-ms p95 latency for one case, with 279 MiB peak VRAM and an approximately 304-MiB index. Figure 12 shows the scaling curve for the measured hardware setup, and Table 15 gives the raw profile.

Table 15: Measured deployment profile on the full stream. Index size combines exact NumPy buffers with approximate Python-container overhead.

Pool	p50 ms	p95 ms	RAM GiB	Peak VRAM MiB	Index MiB	Fit s
2,000	18.4	18.6	1.80	279	5.9	1.4
8,000	17.9	22.7	1.87	279	23.1	2.8
20,000	18.4	18.8	1.98	279	57.7	4.3
50,000	19.8	20.6	2.20	279	143.0	7.7
100,000	23.7	25.1	2.61	279	283.8	11.3
107,136	23.8	25.0	2.69	279	303.9	11.5

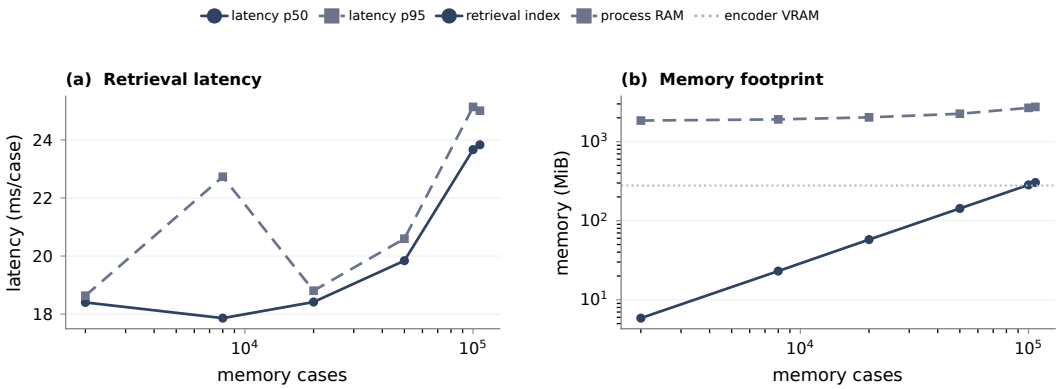


Figure 12: Measured per-case latency (left) and memory footprint (right) as retrieval memory grows. The encoder remains resident on one GPU; physician effects are reported separately in Figure 19.

F MEMORY–AGGREGATION FACTORIAL INTERACTIONS

The main text reports the sequential ledger and Proposition 2. An exploratory 2×2 evaluation crosses frozen versus cumulative evidence with a fixed rule versus LNA+ (Figure 13). The observed interactions are -9.86, -8.44, and -7.81 points: in these cells, changing memory changes the subsequent opportunity available to aggregation. The negative interactions in this factorial quantify path dependence for the tested memory–aggregation combinations.

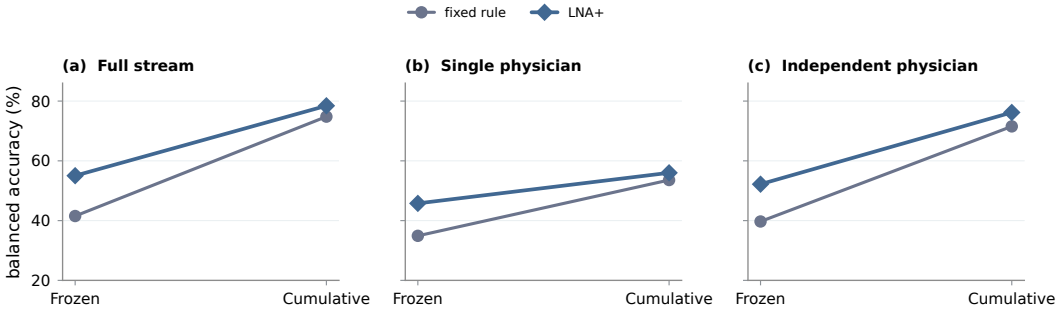


Figure 13: Memory $\times$ aggregation interaction plot. Lines cross frozen and cumulative evidence memory with the fixed rule and LNA+. In the frozen condition the ranker may learn from prior queries but retrieves only warm-up evidence.

Table 16: Conditional simple effects and difference-in-differences interaction in balanced- accuracy points for the exploratory memory–LNA+ factorial.

Stream	Memory effect / hand	Memory effect / LNA+	Aggregation effect / frozen	Aggregation effect / cumulative	Interaction
HOSP-Stream-Full	+33.26	+23.40	+13.49	+3.64	-9.86
HOSP-Stream-SingleMD	+18.67	+10.23	+10.86	+2.42	-8.44
HOSP-Stream-IndependentMD	+31.82	+24.01	+12.48	+4.67	-7.81

G POSITIVE COVERAGE INTERVENTION AND DIAGNOSTIC ROUTING

The removal intervention measures sensitivity to gold-supporting neighbours under a label-conditioned perturbation. Removing neighbours jointly changes support, rank, similarity, and agreement features while holding the fitted model fixed. Figure 14 reports controlled and natural-support results. Expanding lexical and dense neighbourhoods from 32 to 128 increases coverage (Figure 15), but utility depends on how the added evidence is used.

The router's exploratory action set is calibrated memory, externally selected LNA+, and LoRA; it does not enter the confirmatory base-LNA comparison. For each held-out stream, the threshold pair is selected by mean target balanced accuracy on the other two streams over $\tau_C \in \{70, 75, 80, 85, 90, 95\}\%$ and $\tau_H \in \{2, 5, 10, 15, 20, 25\}$ points. The split is stream-object-level only: because Full contains both physician streams, it is not case- or physician-disjoint.

Table 18 exposes every held-out target. Full has 9/12 hindsight matches and SingleMD 6/12; their selected rules reduce to always-LNA+ and always-LoRA respectively. IndependentMD has 1/12 matches: the selected rule assigns all targets to LNA+, while hindsight selects LoRA on 11/12. LOSO selection therefore confirms that the two-signal mapping does not generalize to that stream.

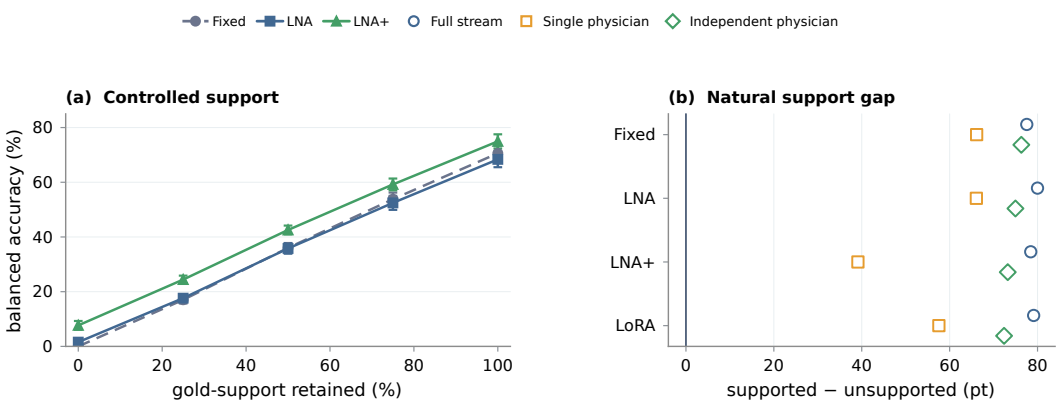


Figure 14: Controlled support intervention. Left: mean balanced accuracy with 95% normal intervals over probe months at each retained-support level. Right: supported-minus- unsupported accuracy under the natural neighbourhoods; marker shape identifies the stream.

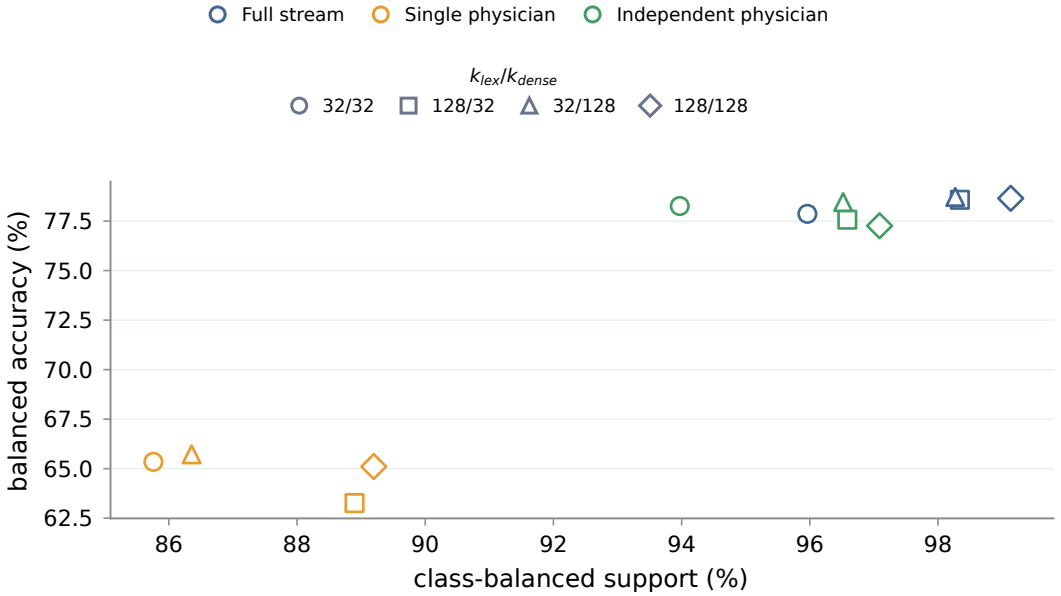


Figure 15: Class-balanced support and LNA+ accuracy for four lexical/dense neighbourhood budgets on each stream. Colour identifies the stream and marker shape the budget pair.

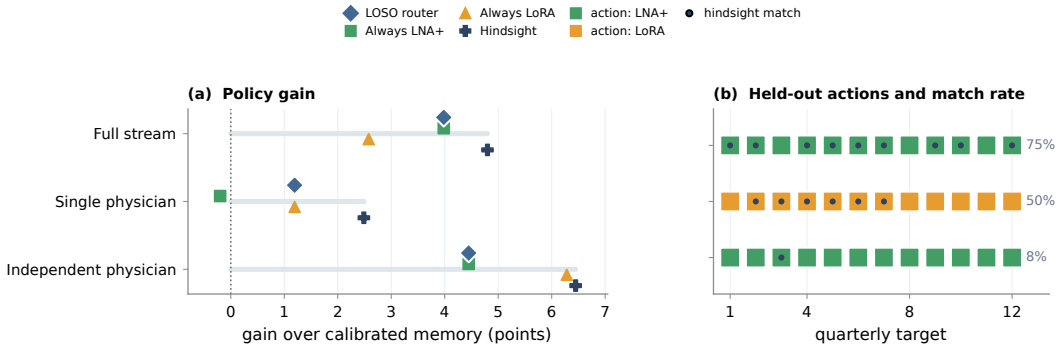


Figure 16: Leave-one-stream-object-out routing. Left: held-out policy gain over calibrated memory, with always-on policies and hindsight comparators. Right: twelve quarterly held-out actions per stream; black centres mark hindsight agreement.

Table 17: Coverage rescue summary. The final column gives the number of cases unsupported at $(32, 32)$ but covered by the expanded budget, followed by their conditional balanced accuracy over represented classes.

Stream	(k_{lex}, k_{dense})	Coverage	Bal. acc.	Rescued cases
HOSP-Stream-Full	(32,32)	95.96	77.86	0 / –
HOSP-Stream-Full	(128,32)	98.34	78.58	33 / 10.00
HOSP-Stream-Full	(32,128)	98.27	78.72	32 / 1.64
HOSP-Stream-Full	(128,128)	99.13	78.65	44 / 7.50
HOSP-Stream-SingleMD	(32,32)	85.76	65.34	0 / –
HOSP-Stream-SingleMD	(128,32)	88.90	63.26	10 / 0.00
HOSP-Stream-SingleMD	(32,128)	86.36	65.72	4 / 0.00
HOSP-Stream-SingleMD	(128,128)	89.20	65.11	12 / 0.00
HOSP-Stream-IndependentMD	(32,32)	93.97	78.26	0 / –
HOSP-Stream-IndependentMD	(128,32)	96.59	77.57	36 / 7.38
HOSP-Stream-IndependentMD	(32,128)	96.52	78.46	35 / 7.14
HOSP-Stream-IndependentMD	(128,128)	97.09	77.26	43 / 8.93

Table 18: Held-out per-probe routing trace with each stream's actual target month. Router and hindsight entries are balanced accuracy (%).

HOSP-Stream-Full				HOSP-Stream-SingleMD				HOSP-Stream-IndependentMD			
Month	Action	Router	Hindsight	Month	Action	Router	Hindsight	Month	Action	Router	Hindsight
2022-07	LNA+	72.45	72.45	2022-07	LoRA	31.33	32.58	2022-08	LNA+	68.15	69.36
2022-10	LNA+	77.45	77.45	2022-10	LoRA	41.59	41.59	2022-11	LNA+	78.03	79.56
2023-01	LNA+	80.07	83.02	2023-01	LoRA	53.84	53.84	2023-02	LNA+	63.89	63.89
2023-04	LNA+	82.04	82.04	2023-04	LoRA	68.53	68.53	2023-05	LNA+	83.05	84.02
2023-07	LNA+	77.97	77.97	2023-07	LoRA	64.34	64.34	2023-08	LNA+	77.51	81.55
2023-10	LNA+	73.54	73.54	2023-10	LoRA	63.96	63.96	2023-11	LNA+	75.46	76.04
2024-01	LNA+	79.52	79.52	2024-01	LoRA	60.79	60.79	2024-02	LNA+	78.72	80.35
2024-04	LNA+	74.55	79.43	2024-04	LoRA	58.53	59.94	2024-05	LNA+	82.10	83.01
2024-07	LNA+	79.99	79.99	2024-07	LoRA	66.86	70.55	2024-08	LNA+	79.50	86.01
2024-10	LNA+	87.03	87.03	2024-10	LoRA	64.88	67.95	2024-11	LNA+	83.45	83.87
2025-01	LNA+	82.52	84.49	2025-01	LoRA	66.51	66.97	2025-02	LNA+	78.00	78.84
2025-04	LNA+	85.53	85.53	2025-04	LoRA	62.40	68.06	2025-05	LNA+	73.80	79.18

H SUPERVISION BUDGET AND TEMPORAL-DEPENDENCE SENSITIVITY

Figure 17 varies the LNA+ fitting budget using nested prefixes of a stratified sample. The reported scores are near 16.67% at 1,000–2,000 neighbourhoods, rise sharply at 3,000, and approach a plateau between 6,000 and 12,000. This curve describes performance as the fitting sample grows. The separate 8,000-case eligible-history threshold controls activation of base LNA and is selected during warm-up; the two quantities serve different roles in the protocol.

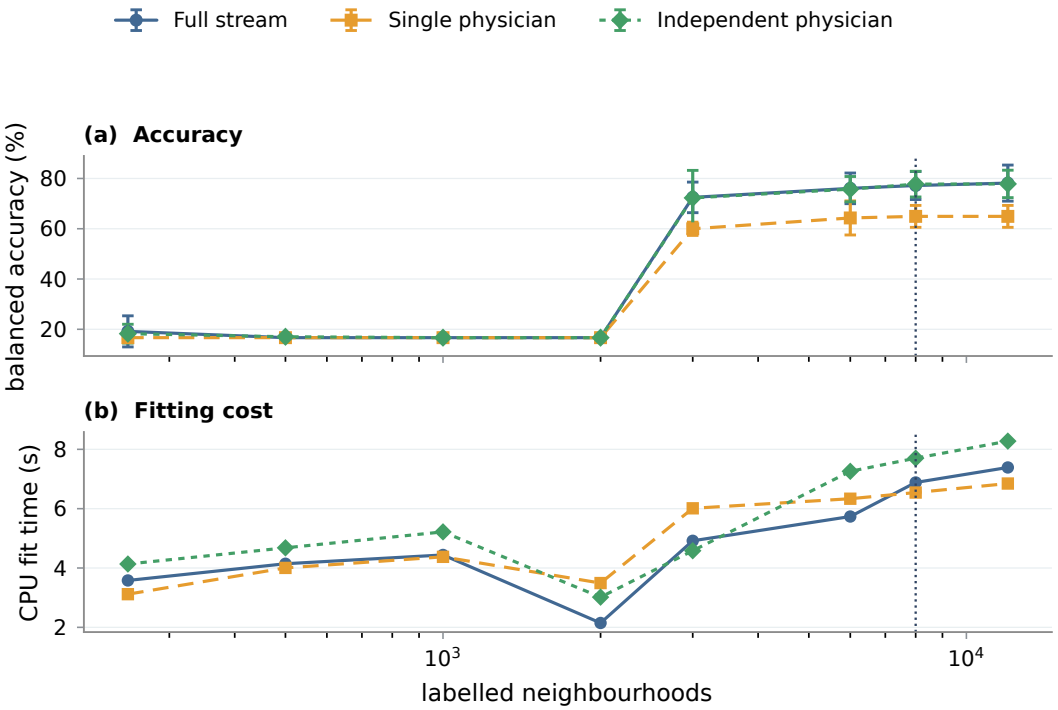


Figure 17: Exploratory LNA+ fitting-budget sensitivity: balanced accuracy (top) and CPU fit time (bottom). Error bars are one standard deviation over every-sixth-month probes. The vertical 8k reference shares the activation threshold's numerical value, but fitting sample size and eligible-history size are different quantities.

For exploratory LNA+, Figure 18 and Table 19 compare IID bootstrap, circular moving-block bootstrap, Newey–West HAC, and quarterly aggregation. Full and IndependentMD gains over calibrated memory remain positive across estimators. SingleMD is sensitive to temporal dependence: its two-sided p-value changes from 0.0054 (IID) to 0.0968 (block-12) and 0.1119 (NW-12). This superiority-sensitivity analysis is specific to LNA+; base-LNA superiority over BL2b retains the primary month-resampling analysis. The separate base-LNA moving-block analysis in Table 14 evaluates non-inferiority to LoRA and retraining.

Table 19: Exploratory LNA+ two-sided p-value sensitivity. Circular blocks have lengths 3, 6, and 12 months; NW-12 uses lag 12. Entries printed as 0.0000 denote rounding at the reported precision.

Stream	Reference	IID	Block-3	Block-6	Block-12	NW-12	Quarterly
HOSP-Stream-Full	calibrated memory	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HOSP-Stream-Full	LoRA	0.0574	0.0416	0.0102	0.0008	0.0004	0.1110
HOSP-Stream-Full	monthly retrain	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HOSP-Stream-SingleMD	calibrated memory	0.0054	0.0132	0.0330	0.0968	0.1119	0.0132
HOSP-Stream-SingleMD	LoRA	0.2786	0.4288	0.5146	0.5886	0.5778	0.4292
HOSP-Stream-SingleMD	monthly retrain	0.9338	0.9676	0.9602	0.9632	0.9781	0.9558
HOSP-Stream-IndependentMD	calibrated memory	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HOSP-Stream-IndependentMD	LoRA	0.0040	0.0082	0.0004	0.0000	0.0000	0.0176
HOSP-Stream-IndependentMD	monthly retrain	0.5716	0.5534	0.5506	0.4788	0.4176	0.5124

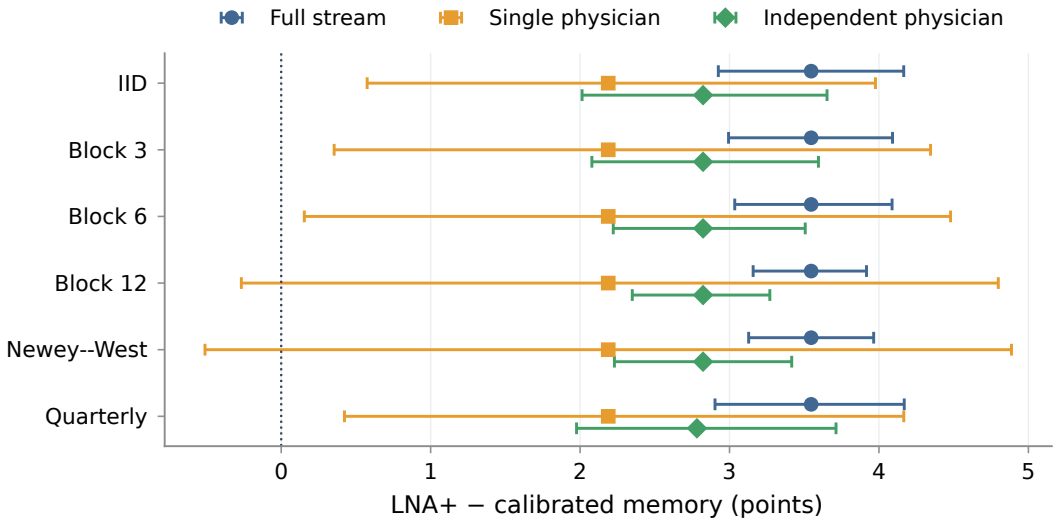


Figure 18: LNA+ minus calibrated-memory effects with 95% intervals under IID, circular-block, Newey–West, and quarterly estimators. Marker shape identifies the stream; the vertical line marks zero effect.

I Physician and class-level sensitivity

Of the 12 recording physicians, four satisfy the outcome-blind eligibility rule: at least 1000 usable cases, 24 observed active months, and 8 scoreable post-warm-up months. The frozen 8k rule then creates two estimands. The deployable policy includes the calibrated fallback before activation; the active-LNA estimand includes only months after the ranker is fitted. Figure 19 reports physician effects with Hartung–Knapp uncertainty and a leave-one-physician-out analysis. Small physician streams remain below 8k; their reported policy differences (+1.77 and -0.27 points) are listed only for the policy estimand, with active-LNA entries left undefined.

With four physicians, interval choice is consequential. Table 21 reports normal random-effects, Hartung–Knapp, modified Hartung–Knapp, a between-physician prediction interval, and the leave-one-physician-out pooled range. The pooled point estimate is +1.73 points, but the Hartung–Knapp interval $[-0.71, +4.17]$ crosses zero and the prediction interval $[-5.16, +8.62]$ spans effects in both directions. Leave-one-physician-out pooled estimates remain positive, while the prediction interval captures substantial between-physician uncertainty.

Class-level localisation (Figure 20) shows that the aggregate gain is not a majority-class artefact. The rare Shaoyang class has the lowest coverage and the largest SingleMD gain, while IndependentMD gains track class-level support-constrained headroom. The pattern profiles class-specific opportunities for broader retrieval and learned decision rules.

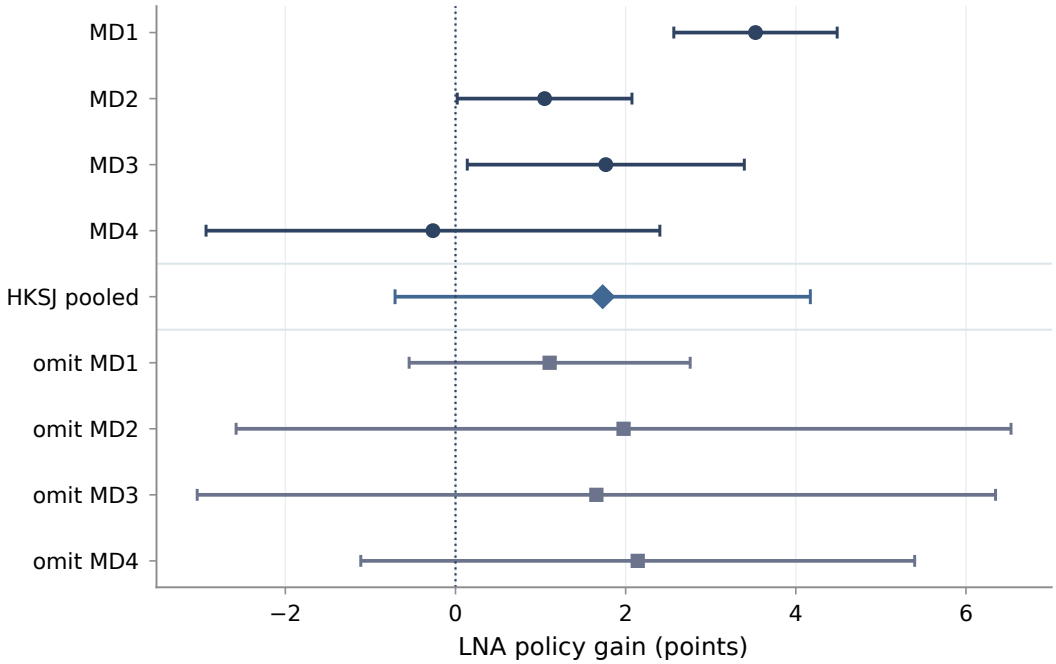


Figure 19: Physician-level policy effects, the Hartung–Knapp pooled estimate, and leave-one-physician-out estimates on a common forest axis. Horizontal intervals are 95% confidence intervals; the vertical line marks zero effect.

Table 20: Mutually exclusive physician streams with separate policy and active-LNA contrasts. Active-LNA entries require at least one month with a fitted ranker. The Hartung–Knapp pooled interval summarizes the policy estimand.

Physician	Cases	Eval. months	Active months	Policy Δ	Active-LNA Δ	Policy 95% CI
`8056e996`	79,535	38	38	+3.53	+3.53	[+2.57, +4.49]
`f3ec4945`	26,099	39	20	+1.05	+1.96	[+0.02, +2.07]
`69d22f35`	6,065	35	0	+1.77	–	[+0.14, +3.39]
`a90628ca`	2,092	11	0	-0.27	–	[-2.93, +2.40]
Random-effects policy	–	–	–	**+1.73**	–	**[-0.71, +4.17]**

Table 21: Physician meta-analysis sensitivity, in balanced-accuracy points. The prediction interval targets the effect in a new eligible physician stream; the leave-one-out row gives the range of pooled point estimates.

Estimator	Pooled effect	Lower	Upper
Normal random-effects CI	+1.73	+0.15	+3.31
Hartung–Knapp CI	+1.73	-0.71	+4.17
Modified Hartung–Knapp CI	+1.73	-0.84	+4.30
Random-effects prediction interval	+1.73	-5.16	+8.62
Leave-one-physician-out range	–	+1.11	+2.14

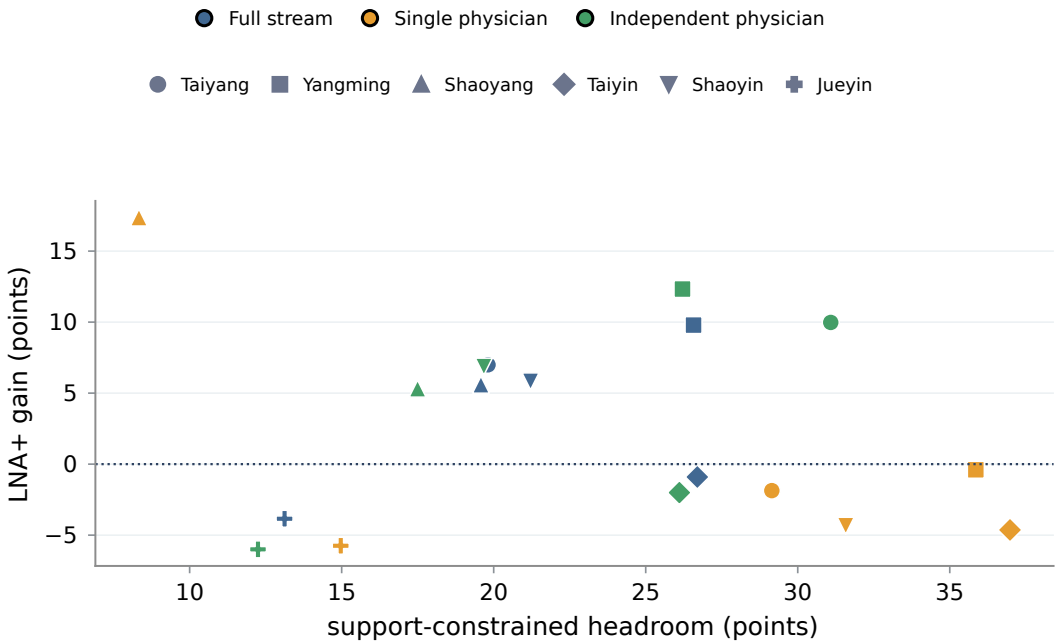


Figure 20: Class-level support-constrained headroom and learned-minus-fixed recall gain. Colour identifies the stream and marker shape the six-meridian class; the horizontal line marks zero gain.

Table 22: Class-balanced retrieval coverage, support-constrained headroom, baseline recall, and LNA+ gain by stream and class.

Stream	Class	Coverage	Headroom	Baseline recall	LNA+ gain
HOSP-Stream-Full	太阳	98.14	19.81	77.16	+6.99
HOSP-Stream-Full	阳明	99.07	26.57	74.13	+9.79
HOSP-Stream-Full	少阳	83.68	19.58	61.77	+5.59
HOSP-Stream-Full	太阴	99.77	26.70	72.85	-0.90
HOSP-Stream-Full	少阴	95.10	21.21	73.89	+5.83
HOSP-Stream-Full	厥阴	100.00	13.12	87.33	-3.85
HOSP-Stream-SingleMD	太阳	96.22	29.15	73.19	-1.86
HOSP-Stream-SingleMD	阳明	97.69	35.86	69.98	-0.40
HOSP-Stream-SingleMD	少阳	18.99	8.33	0.60	+17.37
HOSP-Stream-SingleMD	太阴	99.80	36.98	68.73	-4.63
HOSP-Stream-SingleMD	少阴	65.66	31.58	39.26	-4.34
HOSP-Stream-SingleMD	厥阴	100.00	14.97	88.51	-5.75
HOSP-Stream-IndependentMD	太阳	88.99	31.09	58.70	+9.98
HOSP-Stream-IndependentMD	阳明	97.29	26.21	70.78	+12.33
HOSP-Stream-IndependentMD	少阳	80.75	17.50	60.76	+5.32
HOSP-Stream-IndependentMD	太阴	99.55	26.11	75.28	-2.00
HOSP-Stream-IndependentMD	少阴	95.42	19.68	75.97	+6.86
HOSP-Stream-IndependentMD	厥阴	100.00	12.25	88.44	-6.00

J IMPLEMENTATION AND REPRODUCIBILITY

Table 23 records the exact retrievers, encoder, LNA learners, language model, LoRA configuration, drift trigger and evaluation budget imported from the executable configuration rather than copied manually. The anonymous release contains code, configs, regression tests, result JSON, and table/figure builders, but excludes clinical records and model weights.

Table 23: Implementation manifest. Values are generated from the same configuration objects used by the experiments.

Component	Frozen implementation
Sparse retrieval	IDF-weighted symptom features; $k = 32$; canonical ties
Dense retrieval	BAAI/bge-small-zh; CLS; 512d fp32; max length 128; $k = 32$
LNA base	HGB; 3000 neighbourhoods/month; 200 trees; lr=0.08; 31 leaves
LNA+	HGB; 12000 neighbourhoods/month; positive weight sqrt(C-1); lexical residual 0.5; prior correction 0.25
Language model	Qwen3.5-4B; option-index logits; order seeds 0,1,2
LoRA	rank 16; alpha 32; dropout 0.05; lr=0.0001; 60 steps; batch 1 x accum 16
LoRA targets	q_proj, k_proj, v_proj, o_proj; replay 0.3; 512 examples/update
Drift trigger	input feature PSI, prediction KL, confidence z-score; warm-up FPR 0.2; any 1 fires
Evaluation	12-month warm-up; 200 cases/month; 10000 bootstrap replicates

Experimental estimates are reported from result artefacts; differences calculated from displayed table entries are explicitly marked approximate. The release's checker binds registered headline claims to JSON paths. Protocol mismatches fail the build; retrieval digests must agree under five Python hash seeds, and incremental dense embeddings must equal a one-shot build bit for bit. The release includes the result artefacts and builders needed to regenerate all reported tables and figures.