Supplementary Material: A Label-Efficient On-Device QC Agent for MRI Segmentation under Domain Shift

Jacky Li[0009−0009−3875−2770]

California State Polytechnic University, Pomona, Pomona, CA, USA
jpli@cpp.edu

This supplement provides additional results supporting the main paper: *robustness analyses* of the online update (Sec. 1); *edge-footprint and calibration* measurements (Sec. 2); *clinical review-workload* analyses, including patient-level residual risk (Sec. 3); *cold-start* and *qualitative* evidence (Secs. 4–6); *implementation details* – the 29-dim agent signature and robustness to imperfect reviewer labels (Secs. 7–10); a *replication* on an independent tumor type (Sec. 11); *cross-device* latency (Sec. 12); and a cross-cohort characterization of the entropy anti-ranking phenomenon (Sec. 18), and a predictive empirical trend relating the memory gain to controller strength (Sec. 20). These expand the brief results in the main paper (Tables S1–S22, Figs. S1–S6) with full tables and per-configuration detail.

1 Robustness Analyses

The main-paper online-update result averages over calibration draws and arrival orders. Here we isolate each source of variation in turn – arrival order, fusion hyperparameters, review budget, and the choice of static uncertainty control – to show that the reported gains are not artifacts of a particular configuration.

Implementation details. All experiments use a single fixed configuration. The class-balanced memory risk is the density ratio $r_{\text{mem}}(x) = \rho^1(x)/(\rho^1(x)+\rho^0(x))$ with per-class density $\rho^c(x) = \frac{1}{k}\sum_{i=1}^{k} 1/(\|x-x_i^c\|_2+\varepsilon)$, k=15 nearest neighbours per class, ε=10^{-3}, over the z-scored 29-dim signature (statistics fit on calibration only). The controller is an ℓ_2-regularised logistic regression (C=1.0) on the same signature. The two scores are fused as $r(x) = \frac{1}{2}z(r_{\text{mem}}) + \frac{1}{2}z(r_{\text{ctrl}})$, where each $z(\cdot)$ uses calibration-only 1st/99th percentiles, so the fusion never sees future-stream statistics. Online updating appends each reviewed case's slices and refits the controller with our NumPy solver. The bounded deployment experiment retains the immutable seed plus 2,048 class-balanced recent replay rows. Nothing is tuned on Africa.

Frozen-baseline settings key. The frozen no-memory baseline appears at different values across tables because the *protocol* differs, not the method. Table S1 maps each quoted value to the protocol that produces it. Each number is the baseline *for its own row's protocol*; they are not competing estimates of one quantity.

Table S1. Protocol → frozen-baseline value. The frozen no-memory controller is a single method; its reported AUROC depends on the calibration draw, the cohort mix, and whether the row is a no-memory or a fixed-fusion operating point.

Where reported	Protocol	AUROC
Main Table 1	seed-avg. over ten 5+5 PEDs/GLI draws, external Africa	0.896
Table S2	single fixed calibration split (order-invariant)	0.850
Fusion ablation (main)	controller under shared-normalisation setup	0.880
Table S5	*fixed-fusion* operating point, single split	0.687
Table S4	broad PEDs/GLI/Africa, different cap and cohort mix	varies

Table S2. Stream-order sensitivity on external Africa (cap-5 calibration). The calibration split is fixed; only the online arrival order varies across twenty seeds. The frozen controller is order-invariant.

Method	AUROC	AUPRC	R@20%	R@30%
No-memory, frozen	0.8504	0.6288	0.4054	0.6351
Online controller	0.9468 ± 0.0198	0.8450 ± 0.0820	0.6068 ± 0.0413	0.8166 ± 0.0315
Online memory only	0.9261 ± 0.0077	0.8721 ± 0.0129	0.6057 ± 0.0246	0.7627 ± 0.0186
Fully-online fusion	$\mathbf{0.9545 \pm 0.0114}$	$\mathbf{0.9123 \pm 0.0235}$	$\mathbf{0.6243 \pm 0.0289}$	$\mathbf{0.8226 \pm 0.0328}$

1.1 Stream-Order Robustness

Because the online update appends reviewed cases sequentially, its result could in principle depend on the order in which cases arrive. To isolate this effect from the calibration split, we hold the calibration split fixed (a single seed) and vary only the online arrival order across twenty order seeds. Table S2 reports the resulting spread. The frozen no-memory controller is order-invariant and shown for reference. The online methods are stable: fully-online fusion stays at 0.9545 ± 0.0114 AUROC and 0.9123 ± 0.0235 AUPRC, so its improvement over the frozen baseline is far larger than the order-induced spread and is not an artifact of a favourable arrival order. Point values differ slightly from the main-paper online-update table, which additionally averages over calibration splits.

Fusion weight and memory-k. The fusion combines the memory and controller scores with a fixed 0.5/0.5 weight and k=15 neighbours. Table S3 sweeps the fusion weight over $\{0.3, \ldots, 0.7\}$ and k over $\{5, 15, 30, 50\}$ on external Africa with the calibration split fixed. External-Africa AUROC varies by only 0.012 across the twenty configurations (0.894–0.906), and every configuration stays well above the no-memory controller (0.850); the default is not a tuned operating point.

Review-budget (cap) sensitivity. Table S4 varies the per-dataset review budget across cap 5/10/20 for all three segmenters. The memory gain over the no-memory controller is stable for the spiking segmenter, and grows with the review budget for the weak-confidence segmenters: the SwinUNETR AUPRC gain rises from +0.0088 at cap 5 to +0.0205 at cap 20, and the nnU-Net gain rises from +0.0852 to +0.1871. In every configuration the confidence-aggregation baseline

Table S3. Fusion-weight and memory-k sensitivity on external Africa (AUROC, fixed calibration split). The default configuration is k=15, weight 0.5.

k \ weight	0.3	0.4	0.5	0.6	0.7
5	0.8981	0.8988	0.8992	0.8992	0.8989
15	0.8940	0.8953	0.8960	0.8958	0.8948
30	0.8946	0.8962	0.8976	0.8981	0.8975
50	0.8997	0.9021	0.9040	0.9052	0.9058

Table S4. Review-budget (cap) sensitivity on broad PEDs/GLI/Africa selective review. For each segmenter the memory gain over the no-memory controller is reported at caps 5/10/20; "Review for 80% recall" is the fraction of slices that must be reviewed to catch 80% of major failures.

Segmenter	Cap	Conf.-agg. AUPRC	No-memory AUPRC	Memory AUPRC	Δ AUPRC	Memory R@20%	Review for 80% recall
Spiking	5	0.5292 ± 0.0168	0.9007 ± 0.0103	0.9047 ± 0.0082	+0.0039	0.3563 ± 0.0091	0.5250 ± 0.0178
Spiking	10	0.5323 ± 0.0179	0.9155 ± 0.0112	0.9197 ± 0.0092	+0.0042	0.3599 ± 0.0104	0.5056 ± 0.0244
Spiking	20	0.5326 ± 0.0174	0.9227 ± 0.0059	0.9266 ± 0.0050	+0.0039	0.3606 ± 0.0090	0.4965 ± 0.0151
SwinUNETR	5	0.4676 ± 0.0284	0.7093 ± 0.0411	0.7181 ± 0.0508	+0.0088	0.4531 ± 0.0244	0.5598 ± 0.1194
SwinUNETR	10	0.4759 ± 0.0337	0.7509 ± 0.0284	0.7721 ± 0.0237	+0.0212	0.4728 ± 0.0259	0.4671 ± 0.0569
SwinUNETR	20	0.4824 ± 0.0333	0.7814 ± 0.0357	0.8019 ± 0.0275	+0.0205	0.4859 ± 0.0218	0.4308 ± 0.0273
nnU-Net	5	0.1633 ± 0.0213	0.4081 ± 0.0637	0.4934 ± 0.0760	+0.0852	0.7473 ± 0.0665	0.2451 ± 0.0590
nnU-Net	10	0.1633 ± 0.0213	0.3737 ± 0.0772	0.5413 ± 0.0774	+0.1677	0.7627 ± 0.0733	0.2420 ± 0.0871
nnU-Net	20	0.1633 ± 0.0213	0.4384 ± 0.0930	0.6255 ± 0.0456	+0.1871	0.7937 ± 0.0548	0.2191 ± 0.0675

stays far below both memory and no-memory controllers, and the fraction of slices that must be reviewed to recover 80% of major failures decreases as memory is added and as the cap increases. The effect is therefore not an artifact of one review budget.

Static uncertainty controls. We also ran MC-dropout and ensemble uncertainty controls for the frozen multisite SwinUNETR. MC-dropout hard-mask pairwise disagreement is effectively degenerate in this setting: pairwise-Dice AUROC is 0.500 on both broad and external-Africa audits, indicating that the stochastic passes rarely change the final argmax mask. A true independently trained SwinUNETR ensemble with three seeds is a much stronger control. Table S5 shows that static pairwise-Dice disagreement remains below the learned controller and calibrated fusion on the broad stream, but is the strongest static control on external Africa. This does not contradict the memory-agent claim: the online memory table can be updated after review, and the online memory score beats the static true-ensemble disagreement on external Africa in 9/10 order seeds for AUROC, AUPRC, and R@20%.

2 Edge Footprint and Calibration

Beyond the aggregate footprint quoted in the main paper, we report how the scoring layer's cost grows with the stored memory, how well the fused risk score is calibrated as a probability, and its per-operation CPU budget.

Table S5. SwinUNETR true-ensemble uncertainty controls at cap 5. The ensemble uses three independently trained SwinUNETR seeds. Static pairwise-Dice is a strong external-Africa baseline, while online memory is the updateable deployment-time mechanism. MC-dropout hard-mask disagreement is omitted because its pairwise-Dice AUROC is 0.500 on both settings.

Setting	Method	AUROC	AUPRC	*R*@20%	*R*@30%
Broad PEDs/GLI/Africa	Pairwise-Dice disagreement	0.7435 ± 0.0108	0.7191 ± 0.0118	0.2941 ± 0.0059	0.4397 ± 0.0084
Broad PEDs/GLI/Africa	No-memory controller	0.8498 ± 0.0187	0.8391 ± 0.0183	0.3650 ± 0.0104	0.5289 ± 0.0154
Broad PEDs/GLI/Africa	Calibrated fixed fusion	$\mathbf{0.8584 \pm 0.0201}$	$\mathbf{0.8522 \pm 0.0195}$	$\mathbf{0.3742 \pm 0.0122}$	$\mathbf{0.5387 \pm 0.0162}$
External Africa	Static pairwise-Dice	0.7503 ± 0.0000	0.4384 ± 0.0000	0.4749 ± 0.0000	0.6648 ± 0.0000
External Africa	Fixed fusion	0.6870 ± 0.1344	0.3175 ± 0.1343	0.3224 ± 0.1773	0.4693 ± 0.2058
External Africa	Online memory only	$\mathbf{0.8308 \pm 0.0494}$	$\mathbf{0.5445 \pm 0.1216}$	$\mathbf{0.6000 \pm 0.0867}$	$\mathbf{0.7285 \pm 0.0714}$

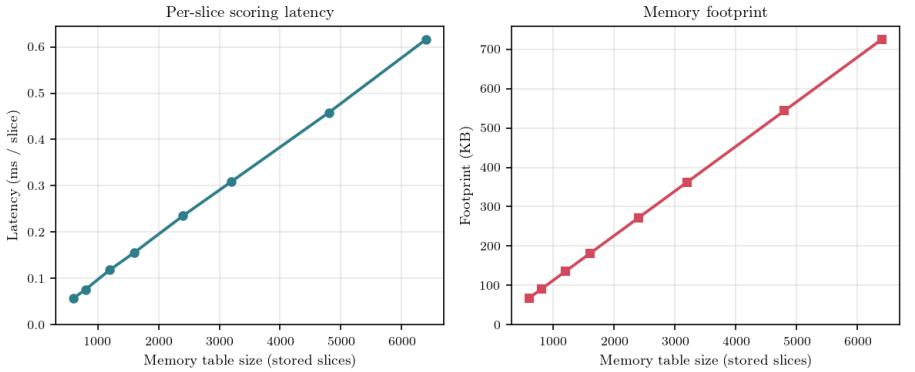

Fig. S1. Edge scaling of the QC layer. Per-slice scoring latency (left) and memory footprint (right) grow linearly with the number of stored slices and stay small, supporting on-device growth of the reviewed-case memory.

Scaling with stored memory. Figure S1 tracks per-slice scoring latency and memory footprint as the reviewed-case table grows. Both scale linearly and remain sub-millisecond and under 1 MB up to several thousand stored slices, so a deployment can keep appending reviewed cases over its lifetime without exceeding the edge budget.

Risk calibration. Because the fused score is used to triage rather than only to rank, its calibration matters. Figure S2 shows reliability curves on external Africa: memory fusion sits closer to the diagonal and lowers expected calibration error from 0.249 to 0.213 relative to the no-memory controller. Absolute calibration on this external cohort remains imperfect, and domain-robust recalibration is left to future work.

Per-operation CPU cost. Table S6 breaks the layer's cost into its three primitives – a class-balanced memory query, the fusion score, and an online append. Each stays below 4.5×10^{-5} s per slice on a single CPU core, confirming that appending reviewed cases is no more expensive than scoring them.

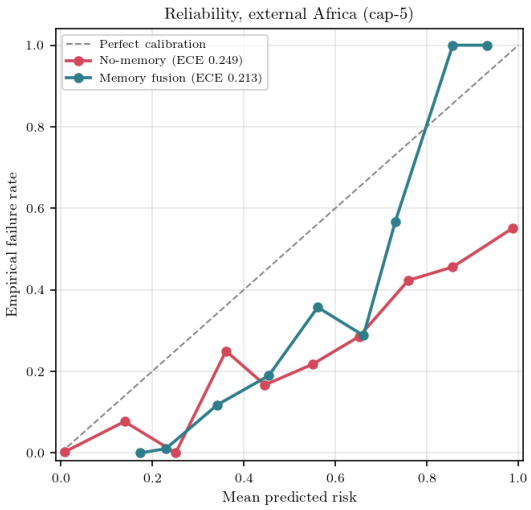

Fig. S2. Reliability on external Africa (cap-5). Memory fusion is closer to the diagonal and has lower ECE than the no-memory controller.

Table S6. CPU memory-query and update footprint for cap-5 external Africa.

Operation	Total seconds	Seconds/slice	Slices/second
Fixed class-balanced query	0.035774 ± 0.001131	0.00003549	28177.12
Fixed fusion score	0.035054 ± 0.000256	0.00003478	28755.26
Online case update	0.044836 ± 0.003421	0.00004448	22482.08

3 Clinical Review Workload Details

We characterise the layer as a triage tool from three angles: how much review it saves at a target failure recall, its clinical net benefit under decision-curve analysis, and – most conservatively – how much residual risk survives at the patient level. Figure S3 reports selective-review efficiency and the decision curve; Fig. S4 gives the per-segmenter selective-review curves; and Tables S7–S8 quantify slice- and patient-level safety.

4 Cold-Start Adaptation

A practical concern for an online method is whether it underperforms before any reviewed cases have accumulated. Figure S5 plots the cold-start learning curve: even at zero reviewed cases the calibration-only fusion already exceeds the frozen no-memory controller, and it improves monotonically as cases are appended – so there is no cold-start penalty to pay for the online mechanism.

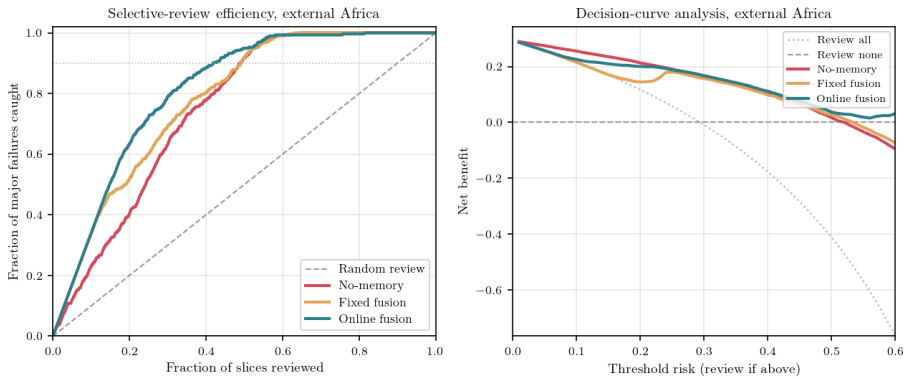

Fig. S3. External Africa (seed-42 split). Left: selective-review efficiency – fraction of major failures caught versus fraction of slices reviewed. Right: decision-curve analysis – clinical net benefit versus the risk threshold at which a clinician chooses to review. Online adaptation reaches any recall target with far less review than the frozen controller and has higher net benefit than reviewing all or none across clinically-relevant thresholds; fully-online fusion is marginally best.

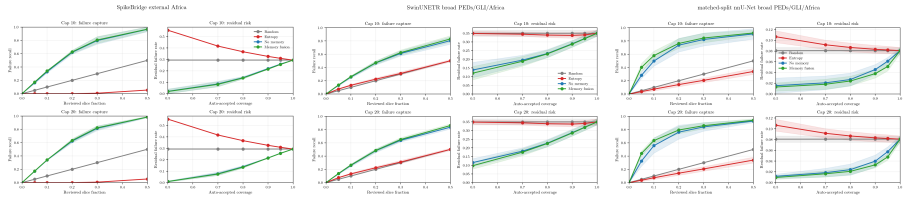

Fig. S4. Selective-review curves for fixed-memory external review and cross-segmenter transfer. Left: spiking segmenter, external Africa. Middle: SwinUNETR broad PEDs/GLI/Africa. Right: matched-split nnU-Net broad PEDs/GLI/Africa.

5 Why Uncertainty Fails: Multi-Cohort Qualitative Panel

Figure S6 is the qualitative motivation referenced from the main paper. It is reproduced here at full width, where the near-zero uncertainty maps on the two confident misses are actually legible.

6 Qualitative Single-Case Zoom

Finally, Fig. S7 illustrates the failure mode that motivates the whole approach: a confidently missed tumor. The frozen segmenter predicts an empty mask while its uncertainty map stays near zero, so confidence-based QC ranks the slice as safe; the memory layer, having seen similar reviewed failures, instead ranks it in the top 5% for review.

Table S7. Review workload and de-prioritised-pool risk on external Africa (cap-5). Left: fraction of slices that must be reviewed to reach a failure-recall target (lower is less clinician effort). Right: residual major-failure rate in the de-prioritised pool at a fixed review budget (lower is safer). Fixed calibration split, matching the main paper's clinical-utility figures: the agent's workload advantage is small (52.9%→51.7% of slices to reach 90% recall) and roughly a third of failures remain in the de-prioritised pool at a 10% budget.

	Review for recall			Residual risk @ budget		
Method	80%	90%	95%	10%	20%	30%
No-memory, frozen	42.2%	52.9%	62.5%	33.7%	26.6%	20.5%
Fixed fusion	42.1%	52.5%	61.7%	33.1%	25.9%	20.5%
Online fusion	**41.6%**	**51.7%**	**58.7%**	**33.1%**	**25.3%**	**19.6%**

Table S8. Patient-level residual risk on external Africa (cap-5): fraction of failing cases that retain at least one de-prioritised (un-reviewed) major-failure slice at each review budget, seed-averaged over ten order seeds. Unlike the slice-level residual risk in Table S6, this strict per-patient criterion stays high at small budgets: slice ranking prioritizes failures well but does not clear an entire patient's failures within a small review budget, underscoring that the layer is a triage aid rather than an autonomous accept/reject decision.

Method	@10%	@20%	@30%
No-memory, frozen	96.8%	90.3%	83.9%
Online controller	96.5%	94.2%	85.3%
Fully-online fusion	98.9%	92.4%	81.1%

7 Slice Signature and Online Update Details

The 29-dimensional signature. Each tumor-relevant slice is summarised by a fixed 29-dimensional feature vector $x \in \mathbb{R}^{29}$ built only from the frozen segmenter's outputs (Table S9). The eight temporal-drift dimensions capture how the prediction changes under a *backbone-specific* perturbation: for the spiking segmenter, two inference budgets (T=2 vs T=4 refinement steps, with entropy/confidence taken from the T=4 softmax); for SwinUNETR, test-time flip augmentation. For the single-pass nnU-Net these dimensions are unavailable and set to zero, so on nnU-Net the signature reduces to its 18 static dimensions – notably, the memory's one resolvable nnU-Net gain therefore comes from the class-balanced density over those static features, not from temporal drift.

Class-balanced memory risk. The memory risk $r_{\text{mem}}(x)$ is a class-balanced inverse-distance density ratio in the standardised 29-dim signature space. Let $d_1^+, \ldots, d_k^+$ and $d_1^-, \ldots, d_k^-$ be the Euclidean distances from x to its k=15 nearest *failure*-class and *success*-class memory entries. With a per-class density $\rho^c(x) = \frac{1}{k}\sum_{i=1}^{k}\left(d_i^c + \varepsilon\right)^{-1}$ (ε=10^{-3}),

$$r_{\text{mem}}(x) = \frac{\rho^+(x)}{\rho^+(x) + \rho^-(x) + 10^{-6}}\,, \tag{1}$$

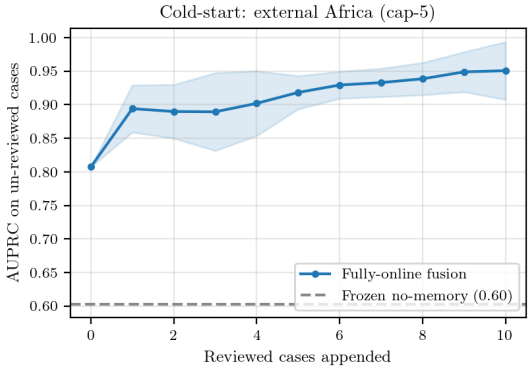

Fig. S5. Cold-start learning curve on external Africa (cap-5). For each of ten arrival orders we append reviewed cases one at a time and evaluate fully-online fusion AUPRC on the *remaining* un-reviewed cases (leakage-clean); the band is ±1 standard deviation over orders. Even at zero reviewed cases the calibration-only fusion is well above the frozen no-memory controller (dashed), and performance climbs steadily as reviewed cases accumulate – there is no cold-start penalty relative to the frozen operating point.

which is near 1 when x lies in a dense region of failures and near 0 among successes. Balancing the two class densities (rather than a single unconditional k-NN distance) prevents the abundant success class from dominating the score when failures are rare. The final risk averages the calibration-percentile normalisations of r_{mem} and the controller.

Online update cadence and deployable solver. In the online variant, cases arrive as a stream. Each case is scored *before* its labels are revealed (leakage-clean); it is then appended to the class-balanced memory table, and the shallow logistic controller is refit on the calibration set plus all reviewed slices accumulated so far. No segmenter gradients are taken and the 29-dim feature extractor is never retrained; the only online state is the memory table, controller history, and

Table S9. The 29-dimensional per-slice signature x (main paper, Slice Signatures). Slice aggregations are mean, top-10% mean, max, and brain-masked mean. All features derive from the frozen segmenter's prediction, softmax, and its change between two temporal inference budgets.

Group	Features	Dims
Prediction context	brain-tissue fraction of the slice	1
Uncertainty	{normalized entropy, 1−confidence} × {mean, top-10%, max, brain-mean}	8
Temporal drift	{softmax Δ, argmax-flip} × {mean, top-10%, max, brain-mean}	8
	patch-level argmax instability × {mean, top-10%, max}	3
Predicted extent	{whole-tumor, tumor-core, enhancing} area fraction	3
	{whole-tumor, tumor-core, enhancing} area jump vs. adjacent slices	3
Component structure	# connected components in {whole-tumor, tumor-core, enhancing}	3
Total		29

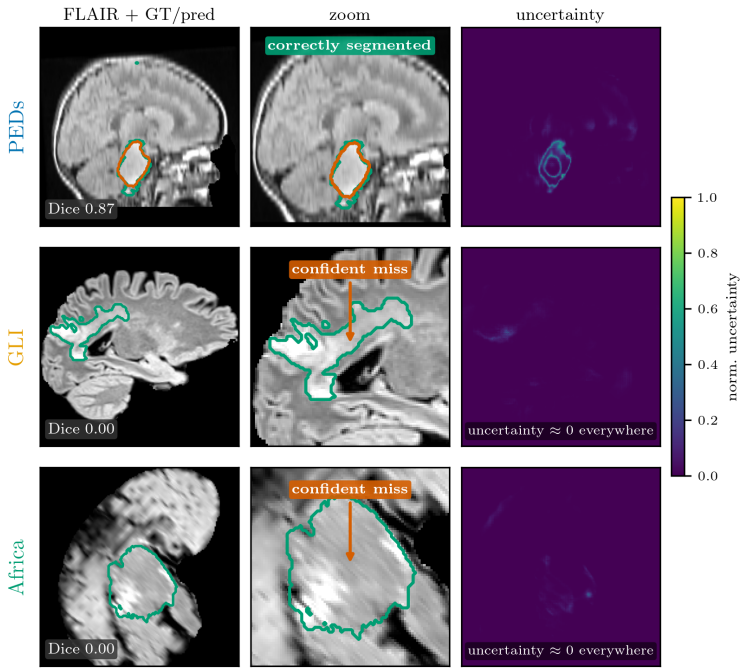

Fig. S6. Why uncertainty fails. Rows: PEDs/GLI/Africa; columns: FLAIR with ground-truth (green) and predicted (orange) contours, zoom, and model uncertainty. On PEDs the segmenter succeeds; on GLI and Africa it misses the tumor entirely (Dice 0.00) yet its uncertainty stays near zero – confident failures that confidence-based QC cannot flag, motivating an agent that learns from reviewed cases.

logistic weights. The runtime we release implements the class-balanced C=1 objective with damped Newton steps in NumPy. Against liblinear its calibration probabilities differ by at most 4.3×10^{-5} (mean 4.8×10^{-6}; rank correlation > 0.999999999), and the calibration-frozen prequential path exactly reproduces the ten-seed 0.9478/0.9311 AUROC/AUPRC result. The guard retains the immutable seed and a deterministic, class-balanced recent coreset in each class.

8 Robustness to Imperfect Review Labels

The online update appends each reviewed slice with a binary major-failure label. In our experiments this label is derived from ground-truth Dice; in deployment it would instead come from a human reviewer and be noisier. To probe this, we randomly flip a fraction of the *appended* labels (the labels used to score AUROC/AUPRC are kept clean) and re-run fully-online fusion, seed-averaged over the same ten seeds. Table S10 shows the degradation is gentle: even at 20% label noise, external-Africa AUROC falls only from 0.948 to 0.917 and stays

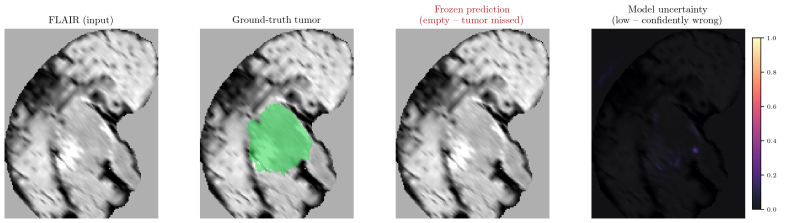

Fig. S7. A confident missed-tumor failure (external Africa, BraTS-SSA-00157, slice 77). Left to right: FLAIR input; ground-truth tumor; the frozen segmenter's prediction (empty – the tumor is missed); and the model's uncertainty map (near zero – the error is made confidently). Confidence-based QC cannot flag this slice, but memory fusion ranks it in the top 5% for review.

Table S10. Robustness of fully-online fusion to noisy appended labels on external Africa (cap-5), seed-averaged over ten seeds. A fraction of appended labels is randomly flipped to simulate imperfect human review; scoring labels are clean. Frozen no-memory baseline (label-noise-free): AUROC 0.896 / AUPRC 0.863.

Appended-label noise	AUROC	AUPRC
0% (oracle Dice labels)	0.9478 ± 0.0042	0.9311 ± 0.0043
5% flips	0.9380 ± 0.0043	0.9191 ± 0.0046
10% flips	0.9312 ± 0.0049	0.9114 ± 0.0051
20% flips	0.9167 ± 0.0072	0.8918 ± 0.0090

far above the frozen no-memory baseline (0.896), so the online-adaptation gain survives realistically imperfect review.

9 Sensitivity to the Major-Failure Dice Threshold

The major-failure label is tumor-relevant $\wedge$ ($\overline{\text{Dice}} < \tau \vee$ missed $\vee$ false) with τ=0.25. Because τ is an operational proxy that ranks failures by overlap rather than clinical severity, and because base failure rates across our cells span 0.045–0.836, the constant is doing substantial work. We therefore re-derive the label at $\tau \in \{0.15, 0.20, 0.25, 0.30, 0.40\}$ and re-run the identical external-Africa protocol at each (ten paper seeds; only the Dice disjunct moves, the missed/false-tumor disjuncts are part of the definition at every τ).

Table S11 shows the online-adaptation gain is *positive at every threshold*, ranging +0.028 to +0.069 AUROC as τ sweeps the base rate from 0.310 to 0.581. The published setting reproduces exactly (0.9478/0.9311 at τ=0.25). The gain is smallest at the strictest threshold, where failures are rarest and the frozen controller is already strong, and largest in the middle of the range; it does not depend on the particular choice τ=0.25.

Table S11. Major-failure Dice-threshold sensitivity on external Africa, seed-averaged over ten seeds. Only the Dice component of the label varies. τ=0.25 is the published setting and reproduces the main-paper values.

τ	base rate	frozen AUROC/AUPRC	online AUROC/AUPRC	ΔAUROC	ΔAUPRC
0.15	0.310	0.9360/0.8753	0.9635/0.9388	+0.028	+0.063
0.20	0.360	0.9088/0.8564	0.9493/0.9276	+0.041	+0.071
0.25	0.398	0.8961/0.8625	**0.9478/0.9311**	+0.052	+0.069
0.30	0.469	0.8575/0.8363	0.9268/0.9247	+0.069	+0.088
0.40	0.581	0.8297/0.8721	0.8939/0.9250	+0.064	+0.053

10 Replay Capacity and Safety-Control Audit

The main paper identifies replay, rather than the auxiliary nearest-neighbour score, as the meaningful memory mechanism. A 512-row buffer is too aggressive: on pediatric LOCO it trails full history by 0.052 AUROC. Increasing the mutable cap to 2,048 rows reduces this gap to 0.012 while retaining resolved gains over last-case-only adaptation (+0.099 AUROC and +0.082 AUPRC). On glioma the corresponding gains are +0.025/ + 0.021; on Africa the stream fits within the cap and bounded replay exactly equals full history. Active state is 0.592 MB and a complete prior checkpoint brings total state to 1.184 MB.

Table S12. Causal replay-memory ablation under LOCO (moved here from the main paper). Deltas compare a 2,048-row bounded replay buffer with last-case-only adaptation or unbounded full history; brackets are paired seed/patient 95% bootstrap intervals.

Held-out cohort	ΔAUROC vs last case	ΔAUPRC vs last case	ΔAUROC vs full history
Africa	$+0.049\,[+0.017, +0.139]$	$+0.075\,[+0.019, +0.199]$	$-0.003\,[-0.008, +0.001]$
Pediatric	$+0.109\,[+0.041, +0.183]$	$+0.091\,[+0.031, +0.137]$	$-0.013\,[-0.024, -0.004]$
Glioma	$+0.034\,[+0.008, +0.067]$	$+0.038\,[+0.009, +0.126]$	$-0.008\,[-0.015, -0.003]$

Capacity vs. selection rule: FIFO and window baselines. To isolate whether class-*balancing* or merely retained *capacity* drives the replay gain, we add at matched capacity a plain recency FIFO (2,048 rows, no balancing) and a 256-row window (Table S13). Two findings. (i) *Capacity* matters beyond the size-1 lower bound: the 2,048-row buffer beats the 256-row window on all three folds (pediatric +0.070 AUROC $[+0.029, +0.102]$, glioma $+0.022\,[+0.003, +0.041]$, Africa $+0.030\,[+0.010, +0.069]$) and pooled ($+0.041\,[+0.006, +0.094]$), so *keeping more history* – not merely more than the last case – is what helps. (ii) Class-balancing is *inert* at matched capacity: the balanced buffer is statistically indistinguishable from the plain FIFO on every fold and pooled (ΔAUROC $-0.001\,[-0.005, +0.003]$).

A simple bounded recency buffer therefore suffices; the balancing machinery is unnecessary.

Table S13. FIFO / window baselines at matched capacity (spiking segmenter, LOCO; two-level seed×patient bootstrap, ΔAUROC). Class-balancing is inert vs. a same-capacity FIFO; capacity (2048 vs. 256) is the driver, now resolved on every fold and pooled since Africa's 95-case stream fills the cap.

ΔAUROC	Pediatric	Glioma	Pooled
bounded − FIFO (same cap)	$-0.001\,[-0.006, +0.004]$	$-0.001\,[-0.003, +0.001]$	$-0.001\,[-0.005, +0.003]$
bounded − window-256	$+0.070\,[+0.029, +0.102]$	$+0.022\,[+0.003, +0.041]$	$+0.041\,[+0.006, +0.094]$
FIFO − last-case	$+0.110\,[+0.042, +0.189]$	$+0.034\,[+0.011, +0.067]$	$+0.065\,[+0.014, +0.171]$

We also audited three automatic controls under disjoint calibration, corrupted stream, and held-out test patients in every LOCO fold: a source- or target-sentinel non-inferiority gate, an immutable risk floor/blend, and quarantine of feedback contradicting a 95% or 99% confident immutable prediction. None produced a cohort-consistent improvement under random flips or a targeted three-case burst; all pooled paired intervals included zero, and several controls hurt pediatric performance. We therefore retain versioning and exact rollback only as recovery mechanisms triggered by external monitoring, not as validated automatic safety guards. Full exploratory results and scripts accompany the code release.

11 Replication on an Independent Tumor Type: BraTS-MEN-RT

To test the online-adaptation claim beyond the Africa cohort, we run the within-domain online protocol on BraTS-MEN-RT (meningioma radiotherapy): 8 calibration subjects, 166 test subjects (25,200 slices), base major-failure rate 0.834, seed-averaged over the same ten seeds. This is a genuinely independent cohort – a different tumor type produced by a different segmenter (SegResNet), foreground-only, so the tumor-core/enhancing signature dimensions are degenerate and the high base rate makes AUPRC uninformative (AUROC is the operative metric). Table S14 shows the Africa pattern replicates: online adaptation lifts AUROC from 0.819 (frozen) to 0.878 (fully-online fusion; +0.059, 10/10 seeds) and cuts variance fivefold, with the memory pathway a small but consistent increment over the online controller (+0.004 AUROC, 10/10). Online adaptation is again the dominant effect and memory a marginal refinement – the same conclusion as on Africa, on data that overlaps in neither site, tumor type, nor segmenter.

12 On-Device Latency Across Hardware

Table S15 benchmarks our NumPy-only fixed scorer on a workstation CPU, an embedded clinical-edge accelerator platform (NVIDIA Jetson Orin Nano), and

Table S14. Online adaptation on an independent tumor type (BraTS-MEN-RT), seed-averaged over ten seeds. Within-cohort deployment: 8 calibration subjects, 166 test subjects. AUROC is the operative metric (base rate 0.834 makes AUPRC trivially high for all methods).

Method	AUROC	AUPRC
No-memory, frozen	0.8192 ± 0.0362	0.9452 ± 0.0152
Online controller	0.8740 ± 0.0052	0.9651 ± 0.0021
Online memory only	0.8255 ± 0.0125	0.9501 ± 0.0043
Fully-online fusion	$\mathbf{0.8778 \pm 0.0054}$	$\mathbf{0.9667 \pm 0.0022}$

two GPU-less single-board computers, confirming the layer runs on genuinely constrained hardware. The scorer is deterministic, so scores are reproducible per device and agree across devices up to floating-point tolerance; only latency varies, and even on a Raspberry Pi it is sub-millisecond per slice and negligible relative to 3D segmentation inference. On the Jetson the fixed scorer reaches 90 μs/slice inside a 15 W power envelope. The complete case-stream plus controller-refit path was also measured on this device at 247 μs/slice (4,053 slices/s over 20 repeats); memory append alone was 90 μs/slice.

Table S15. End-to-end per-slice latency of our numpy-only QC scorer across devices (cap-5 external-Africa stream, k=15; not directly comparable to the per-operation reference timings in Table S6). Scores agree across devices up to floating-point tolerance; only speed differs. All rows are CPU-only under each platform's default governor; the scorer is NumPy-only by construction and does not use the Jetson's GPU.

Device	CPU	μs/slice	slices/s
Workstation	x86-64, 20-core	65	15,300
Jetson Orin Nano (15 W)	Cortex-A78AE, 6-core	90	11,100
Raspberry Pi 4	Cortex-A72, 4-core	354	2,820
Raspberry Pi 400	Cortex-A72, 4-core	344	2,900

The Jetson exposes on-board INA3221 rails, so on that platform we also measure power directly. Sampling at 20 Hz, the board draws 3.32 W idle and 5.18 W under sustained scoring, i.e. the marginal power cost of the QC layer is 1.86 W, for 163 μJ per slice of active energy (455 μJ including idle board power). Over a 60 s soak, throughput is flat (drift $< 0.3\%$) and the junction temperature rises only 48.8 $\rightarrow$ 50.1 °C, so the quoted rate is a sustained rather than a burst figure. The Pi rows remain latency-and-footprint only: those boards expose no comparable rails, so we do not claim cross-device power comparisons. Peak resident memory during scoring stays within the <1 MB model plus the transient signature batch on every device tested. The Pi 4 and Pi 400 share the same Cortex-A72 SoC, so those two rows bound one platform rather than two distinct ones.

The historical device latency/power rows time fixed fusion plus memory append only. On the workstation, the corrected case-stream benchmark scores each

of 15 cases before review and then performs a NumPy refit: 62 μs/slice versus 44 μs/slice fixed. The complete case-stream plus controller-refit path was subsequently measured *on the Jetson Orin Nano* at 247 μs/slice (4,053 slices/s, 2.74× its fixed-score path). Full uncapped state after all 1008 Africa slices is 0.360 MB; the 2,048-row capped replay/controller state is 0.592 MB (1.184 MB with one rollback checkpoint).

Segmenter and agent can co-reside. The timings above cover only the QC layer. To test whether the *segmenter* can share the same edge device, we ran two 3D segmenters on the Orin's GPU (MONAI sliding window, overlap = 0.5, gaussian blending) via a no-sudo user-space PyTorch 2.13/CUDA 13 install, on a BraTS-native 240×240×155 four-modality volume. A 3D SegResNet at roi = 128^3 runs end-to-end in 77.9 s (fp16, two windows per batch; 85.8 s fp32), drawing 8.4 W at a peak GPU allocation of only 1.6 GB – comfortably inside the 8 GB unified memory. **SwinUNETR** – one of the segmenters we evaluate – is heavier: at the paper's 128^3 evaluation ROI it *exhausts* the 8 GB (out-of-memory), but at its 96^3 training ROI it fits, running in 158.6 s at 1.7 GB peak and 9.2 W (fp16). Half precision and window-batching cut memory but barely latency ($\leq 9\%$): these small-batch 3D convolutions are launch- and memory-bound on this GPU, not tensor-core-bound. Thus an evaluated segmenter and the adaptive QC agent fit on a single 15 W device with no server and neither images nor feedback leaving it (a fully data-local station for privacy-sensitive or low-resource sites), though a heavier backbone needs a reduced inference ROI, and the trade is throughput (∼78–159 s/volume for segmentation vs 90 μs/slice for fixed QC). This is a co-residence result, not measured end-to-end latency: signature extraction and any extra temporal-feature passes are excluded.

13 Deployment-Realistic Review: Budgeted Labeling and Asymmetric Noise

The noisy-label study above flips appended labels *uniformly at random*. Two further departures from the idealised protocol are closer to real review, and both were flagged as threats to the online-adaptation claim.

Budgeted labeling (risk-selected memory). Our main protocol appends labels for *all* slices of each reviewed case. Under a real review budget, only the highest-risk slices of a case would actually be inspected and labeled, so the memory would grow from a *risk-selected*, non-random subset – a potential selection bias. We re-run fully-online fusion appending, at each step, true labels only for the top-b fraction of the current case's slices *by the agent's own current risk rank* (at least the single highest-risk slice is always kept); de-prioritised slices never enter memory. Scoring labels remain clean and complete. Table S16 shows graceful, monotone degradation: even at a 10% review budget the agent loses only ∼1.4 AUROC points and stays far above the frozen baseline (0.896). Adaptation therefore survives training on the risk-selected subset a review budget would actually produce.

Table S16. Budgeted labeling on external Africa (cap-5), seed-averaged. Only the top-b fraction of each reviewed case's slices, ranked by the agent's current risk, receive appended labels; scoring labels are clean. "full" reproduces the main-paper protocol. Frozen no-memory baseline: AUROC 0.896 / AUPRC 0.863.

Review budget b	AUROC	AUPRC
full (1.00)	0.9478 ± 0.0042	0.9311 ± 0.0043
0.30	0.9312 ± 0.0114	0.9161 ± 0.0117
0.20	0.9193 ± 0.0141	0.9050 ± 0.0162
0.10	0.9043 ± 0.0167	0.8899 ± 0.0201

Table S17. Asymmetric reviewer noise on external Africa (cap-5), seed-averaged. Only true-failure appended labels are flipped to success at rate f; true successes are never flipped. Scoring labels are clean.

Failure-only flip f	AUROC	AUPRC
0.00	0.9478 ± 0.0042	0.9311 ± 0.0043
0.05	0.9478 ± 0.0044	0.9314 ± 0.0048
0.10	0.9477 ± 0.0042	0.9312 ± 0.0050
0.20	0.9465 ± 0.0047	0.9283 ± 0.0071

Asymmetric (class-conditional) reviewer error. Real reviewer error is unlikely to be symmetric: a rushed reviewer is more apt to *miss* a failure (label a true failure as a success) than to invent one. We therefore flip only true-failure appended labels ($1 \to 0$) at rate f, leaving true successes intact. Table S17 shows this is, if anything, milder than the symmetric noise of Table S10 (AUROC 0.947 at f=0.20 vs 0.917 under symmetric 20%): because failures are the minority class, flipping only failures corrupts far fewer labels in absolute terms. The agent is thus robust to the failure-missing error structure most plausible in deployment.

14 Updateability Under Domain Shift: Online Agent vs. a Frozen Static Rule

The agent's claimed durable advantage over a strong static detector is *updateability*. We test this directly with a non-stationary stream. Per seed: (i) *calibration* – cap-5 PEDs and cap-5 GLI subjects, used once to orient and *freeze* the static rule, seed the agent's memory, fit its initial controller, and standardise; (ii) a *pre-shift burn-in* of further PEDs/GLI subjects in shuffled order; then (iii) a *shift* in which the arriving domain switches to external BraTS-Africa. The static detector is the calibration-sign-corrected entropy oriented once and never updated; the online agent is fully-online fusion, whose memory grows and controller refits case-by-case through the entire stream (labels revealed only after each case is scored). We report performance on the post-shift (external-Africa) portion, seed-averaged over ten seeds.

Table S18 shows the online agent exceeds the frozen static rule on every seed post-shift, and on the full 95-subject cohort the margin resolves under the two-level hierarchical bootstrap (ΔAUROC $+0.124\,[+0.087, +0.161]$, ΔAUPRC

Table S18. Updateability under domain shift (post-shift external Africa, ten seeds). The static rule is a calibration-sign-corrected entropy oriented once and frozen; the online agent keeps adapting through the shifting stream.

Method (post-shift)	AUROC	AUPRC
Static sign-corrected entropy (frozen)	0.8250 ± 0.0078	0.7975 ± 0.0057
Online controller	0.9473 ± 0.0049	0.9298 ± 0.0056
Online memory only	0.9028 ± 0.0102	0.8771 ± 0.0104
Fully-online fusion	$\mathbf{0.9489 \pm 0.0039}$	$\mathbf{0.9330 \pm 0.0039}$

Table S19. nnU-Net memory gain against an online controller (broad PEDs/GLI/Africa, cap-20, ten seeds). Online refitting barely moves the no-memory controller; memory fusion adds a large, all-seed-consistent AUPRC increment.

Method	AUROC	AUPRC
No-memory, frozen	0.8742 ± 0.0399	0.4473 ± 0.1175
Online controller (no memory)	0.8849 ± 0.0214	0.4609 ± 0.0719
Online memory only	0.8532 ± 0.0183	0.6060 ± 0.0474
Fully-online fusion	$\mathbf{0.9084 \pm 0.0167}$	$\mathbf{0.6431 \pm 0.0346}$

$+0.136\,[+0.098, +0.175]$, 4000 resamples). Unlike the 15-case subset – where the frozen sign-corrected entropy held $\sim$0.93 AUROC and the agent's edge was modest and unresolved – on the full cohort the static rule degrades under the external shift (to $\sim$0.83 AUROC) while the online agent holds $\sim$0.95, so updateability recovers a resolved edge as the frozen baseline decays.

15 The nnU-Net Memory Gain Survives an Online Controller

The main paper reports that on matched-split nnU-Net the memory resolves a real gain where entropy is near-useless. A simple version of that comparison is against a *static* no-memory controller, so we verify it against an *adapting* one. On the broad PEDs/GLI/Africa split (cap-20 calibration, $\sim$165 test cases/seed, ten seeds) we run the full online protocol and let the no-memory controller refit online on the same reviewed stream. Table S19 shows that online refitting lifts the no-memory controller's AUPRC only from 0.447 to 0.461, whereas adding online memory fusion reaches 0.643 – a +0.182 AUPRC increment over the online controller that wins all ten seeds and is resolved under the two-level hierarchical bootstrap (seeds $\times$ patients, 4000 resamples): ΔAUPRC $+0.182\,[+0.071, +0.311]$, excluding 0. The smaller ΔAUROC +0.024 (9/10 wins) is *not* resolved by that instrument (CI $[-0.008, +0.052]$ includes 0). The memory gain is therefore not an artifact of a frozen baseline, but it is carried by AUPRC, not AUROC. Note that online memory *alone* helps AUPRC but slightly hurts AUROC; it is the fusion that dominates on both metrics.

Table S20. Static-refit control on external Africa. V1 = full-exposure leave-one-subject-out static refit; V2 = 5-subject on-site static refit (tested on the held-out 90). A full static refit slightly beats the online agent; the agent's edge is over the small-budget refit.

Method	Test set	AUROC	AUPRC
Frozen (PEDs/GLI only)	full Africa	0.8961 ± 0.0351	0.8625 ± 0.0434
Static refit V1a (Africa-only LOSO)	full Africa	0.9591	0.9426
Static refit V1b (cal + Africa LOSO)	full Africa	0.9589 ± 0.0007	0.9425 ± 0.0012
Static refit V2 (5-subject on-site)	held-out 90	0.8636 ± 0.0998	0.8072 ± 0.1527
Fully-online fusion (agent)	full Africa	0.9478 ± 0.0042	0.9311 ± 0.0043

16 Is the Gain Just Target-Domain Recalibration? A Static-Refit Control

The frozen baseline is calibrated on PEDs/GLI and tested on Africa, so its deficit could be pure domain mismatch that *any* target-domain refit would fix. We build that missing control: a plain static logistic controller (no memory, no streaming) refit on Africa data, at two exposure levels. **V1 (full exposure)** is leave-one-subject-out over all 95 Africa subjects – each held-out subject is scored by a controller fit on the other 94 (V1a: Africa only; V1b: plus the 5 PEDs + 5 GLI calibration subjects). **V2 (small budget)** fits on the first 5 Africa subjects only and tests on the remaining 90, seed-averaged, with the frozen and online agents re-evaluated on the same held-out 90 for a paired comparison. Frozen and online-fusion numbers reproduce the published values exactly before the control is introduced.

Table S20 shows a full static refit (V1b) reaches 0.959/0.943, edging out the online agent (0.948/0.931; paired online−V1b = −0.011 AUROC, V1b wins 10/10). So the online machinery does *not* exceed a full target-domain recalibration. Its genuine advantage is label-efficiency: it nearly attains full-refit quality from streamed case-by-case labels only, and against a realistic 5-subject on-site calibration (V2) it keeps a consistent margin (+0.087 AUROC, +0.126 AUPRC, 10/10 matched folds). We accordingly frame online adaptation as label-efficient, not as accuracy beyond recalibration.

17 Multi-Cohort External Validation (Leave-One-Cohort-Out)

External Africa now comprises 95 out-of-institution glioma cases, enough to resolve the online-adaptation gain on that cohort alone (Table S21). To further test generalization across populations, we repeat the full online protocol under a leave-one-cohort-out (LOCO) design: hold out the pediatric (n=223), adult-glioma (n=244), and African (n=95) cohorts in turn, calibrating on 5+5 subjects from the other two, and pool the three held-out test sets (≈ 562 patients; the 95 African cases are out-of-institution, the rest cross-cohort BraTS). All CIs are the strict two-level bootstrap (resample seed, then patients within, 4000

resamples); spiking segmenter. CIs are reported per comparison at nominal 95% without family-wise or false-discovery correction across the tables; borderline resolutions (lower bound near 0) should be read accordingly, and we rest the online-adaptation claim on the pooled effect and its consistent per-cohort sign rather than on any single marginal interval.

Table S21 reports the result. **The pooled gain resolves, as does the out-of-institution African cohort alone** (per-cohort intervals uncorrected for multiplicity; we rest the claim on the pooled effect and consistent per-cohort sign). Online adaptation (agent vs. frozen) reaches pooled ΔAUROC $+0.090$ $[+0.054, +0.154]$ with all three cohorts excluding 0 (Africa $+0.052$ $[+0.015, +0.155]$); label-efficiency (agent vs. a 5-subject refit) reaches pooled $+0.104$ $[+0.049, +0.181]$, resolved on every cohort. The online gain scales with frozen-transfer difficulty (pediatric $+0.108$, Africa $+0.052$, glioma $+0.048$), the expected behaviour of a domain-adaptation mechanism. The memory's increment over the online controller no longer resolves pooled and is negligible ($+0.0024$ $[-0.0004, +0.0045]$ AUROC), consistent with the controller carrying the adaptation and the memory's decisive role being confined to nnU-Net.

Table S21. Leave-one-cohort-out external validation (spiking segmenter; two-level bootstrap, 4000 resamples). Pooled rows combine the three held-out test sets (≈ 562 patients). "excl. 0" = 95% CI excludes zero.

Comparison	Cohort	ΔAUROC [95% CI]	excl. 0
Agent vs. frozen	africa (n=95)	$+0.052$ $[+0.015, +0.155]$	yes
	peds (n=223)	$+0.108$ $[+0.059, +0.211]$	yes
	gli (n=244)	$+0.048$ $[+0.014, +0.101]$	yes
	pooled	$\mathbf{+0.090}$ $[\mathbf{+0.054}, \mathbf{+0.154}]$	**yes**
Label-eff. (agent vs. 5-subj refit)	africa (n=95)	$+0.087$ $[+0.010, +0.390]$	yes
	peds (n=223)	$+0.094$ $[+0.031, +0.207]$	yes
	gli (n=244)	$+0.099$ $[+0.029, +0.375]$	yes
	pooled	$\mathbf{+0.104}$ $[\mathbf{+0.049}, \mathbf{+0.181}]$	**yes**
Memory vs. controller pooled		$+0.0024$ $[-0.0004, +0.0045]$	no (negligible)

Closing the loop: routing-gated learning. The main online protocol appends every reviewed slice, so the routing action does not itself gate learning. We test the fully closed loop under LOCO: at each case, append only the top-20% slices the agent *itself* routes for review (by current controller risk), so the action gates what is learned (Table S22). All three LOCO folds are covered, including the out-of-institution African cohort. The gated loop beats frozen on AUPRC in *every* cohort and pooled ($+0.052$ $[+0.003, +0.137]$), and on AUROC in Africa ($+0.028$ $[+0.001, +0.096]$) and pediatric ($+0.085$ $[+0.046, +0.178]$); the pooled AUROC gain is positive with its lower bound at zero ($+0.047$ $[+0.000, +0.159]$), so we do not lean on it. Closing the loop is *not* free: against the open append-all

protocol it costs a resolved pooled AUROC deficit ($-0.019\,[-0.049, -0.002]$), resolved on pediatric and glioma individually. The agent therefore reaches most of the adaptation gain from only the slices it flags — a genuinely closed perceive→act→learn loop at a fifth of the labels, at a small but measurable accuracy cost.

Table S22. Closed-loop (routing-gated) online adaptation under LOCO: append only the top-20% slices the agent routes for review. Two-level seed×patient bootstrap. All three LOCO folds are covered. The gated gain over frozen resolves on AUPRC in every cohort and pooled; the pooled AUROC lower bound sits at zero. Gating is not free: against the open append-all protocol it costs a resolved pooled $-0.019\,[-0.049, -0.002]$ AUROC.

	Africa	Pediatric	Glioma	Pooled
gated − frozen (AUROC)	$+0.028\,[+0.001, +0.096]$	$+0.085\,[+0.046, +0.178]$	$+0.029\,[-0.003, +0.085]$	$+0.047\,[+0.000, +0.159]$
gated − frozen (AUPRC)	$+0.043\,[+0.003, +0.114]$	$+0.071\,[+0.039, +0.140]$	$+0.041\,[+0.001, +0.142]$	$+0.052\,[+0.003, +0.137]$
gated − ungated (AUROC)	$-0.021\,[-0.059, +0.003]$	$-0.018\,[-0.042, -0.001]$	$-0.017\,[-0.029, -0.006]$	$-0.019\,[-0.049, -0.002]$

18 Entropy Anti-Ranking Across Cohorts and Segmenters

The main paper motivates the agent with a single observation on external Africa: slice-mean predictive entropy *anti-ranks* major failures on tumor-relevant slices (AUROC 0.167, i.e. below chance), because a confidently missed tumor is an empty, low-entropy prediction. Table S23 shows this is not an Africa quirk. Across all ten cohort×segmenter cells we test, raw entropy AUROC on the tumor-relevant subset is *below* 0.5 *in* 10/10 *cells* (range 0.16–0.49): entropy always anti-ranks the failures. The *severity* tracks the mechanism – the fraction of failures that are confident misses correlates −0.64 with entropy AUROC – so the effect is dramatic only where confident misses dominate (spiking GLI/Africa, nnU-Net PEDs/Africa) and mild where they are rare (SwinUNETR, SegResNet MEN-RT).

Two honest caveats. (i) The anti-ranking is induced by the tumor-relevant filter, which uses ground truth: on *all* slices, entropy AUROC is above 0.5 everywhere (0.61–0.89), so the phenomenon holds for the tumor-relevant population but is not observable at inference without the label-dependent filter. (ii) A calibration-oriented sign flip recovers a moderately strong detector (0.79–0.86 AUROC) only in the deep-anti-ranking cells; in most cells it reaches an unremarkable 0.47–0.62, so sign-corrected entropy is a diagnostic of the artifact, not a general deployable detector. Sign-corrected AUROC is out-of-sample (orient on a 50% subject split, evaluate on the held-out half, ten seeds).

Does the gain come from recalibrating an entropy scalar online? A natural challenge to our online-adaptation claim is that simply *recalibrating entropy online* – refitting a 1-D logistic on standardized `entropy_mean` over calibration

Table S23. Entropy anti-ranking is universal but severity varies. Per cohort×segmenter, on tumor-relevant slices: raw entropy failure-AUROC (filtered and, for contrast, unfiltered), out-of-sample sign-corrected AUROC, the fraction of failures that are confident misses, and the base failure rate. Raw filtered AUROC < 0.5 in every cell; unfiltered > 0.5 in every cell. The headline LOCO/online results use all 95 SSA-glioma cases (spiking segmenter/Africa here); the other segmenters report each one's precomputed evaluation set.

Segmenter	Cohort	raw AUROC (filtered)	raw AUROC (unfilt.)	sign-corr. AUROC	%conf. miss	base rate
Spiking	PEDs	0.429	0.804	0.573	6%	0.561
Spiking	GLI	0.169	0.679	0.832	57%	0.545
Spiking	Africa	0.167	0.621	0.843	45%	0.398
nnU-Net	PEDs	0.260	0.738	0.794	55%	0.171
nnU-Net	GLI	0.410	0.642	0.582	48%	0.059
nnU-Net	Africa	0.156	0.608	0.864	29%	0.045
SwinUNETR	PEDs	0.492	0.805	0.473	5%	0.627
SwinUNETR	GLI	0.384	0.770	0.618	11%	0.273
SwinUNETR	Africa	0.277	0.657	0.717	12%	0.149
SegResNet	MEN-RT	0.404	0.888	0.598	2%	0.836

+ reviewed-so-far slices, per case arrival, same stream as the online controller, no memory and no 29-dim controller – might capture the same gain far more cheaply. It does not. On external Africa (seed-averaged, Table S24) this online-recalibrated entropy reaches AUROC/AUPRC 0.825/0.783, versus the agent's 0.948/0.931: the agent wins by +0.123/ + 0.148 on all ten seeds, and on the full 95-subject cohort the gap resolves under the two-level hierarchical bootstrap (agent−recal-entropy ΔAUPRC $+0.148\,[+0.110, +0.188]$). Two further checks confirm the direction: (i) recalibrating entropy online does *not* beat *freezing* the sign-corrected entropy (0.783 vs 0.798 AUPRC), so online recalibration of a scalar does not help; and (ii) the strongest frozen entropy variant still trails the agent by >0.13 AUPRC. So the online adaptation is carried by the streamed 29-dim controller and memory interaction; recalibrating a confidence scalar – online or frozen – does not reach the agent. Note this control also subsumes a *periodically re-oriented* sign-corrected entropy: a 1-D logistic on standardised entropy is free to learn a negative coefficient, so re-orienting the sign on the accumulating reviewed stream is a special case of the rule tested here, and it reaches 0.825/0.783.

19 Difficulty-Correlated ("Miss-the-Miss") Reviewer Noise

The noise studies above flip labels at random. The most adversarial and realistic error is a hurried reviewer who fails to flag a *real* failure, preferentially on failures that *look* acceptable – a "miss-the-miss". We therefore flip only true-failure appended labels (1→0) and concentrate the flips, at a fixed overall rate, on the subtlest failures by a per-slice subtlety score $s = \frac{1}{2}\,\mathrm{clip}(\mathrm{Dice}/0.25, 0, 1) + \frac{1}{2}(1 - \mathrm{rank}(\text{lesion size}))$ (near-threshold Dice and small lesions score high); per-slice

Table S24. Online-recalibrated entropy vs. the agent on external Africa (seed-averaged). Recalibrating an entropy scalar online – with or against a frozen entropy baseline – does not close the gap to the controller+memory agent.

Method	AUROC	AUPRC
Online-recalibrated entropy (1-D)	0.825 ± 0.003	0.783 ± 0.006
Frozen sign-corrected entropy	0.827 ± 0.006	0.798 ± 0.004
Online controller (29-dim, no memory)	0.947 ± 0.005	0.929 ± 0.006
Online fusion (agent)	$\mathbf{0.948 \pm 0.004}$	$\mathbf{0.931 \pm 0.004}$

Table S25. Difficulty-correlated reviewer noise on external Africa (cap-5), seed-averaged, vs. the symmetric-uniform and asymmetric-random sweeps. Flips concentrated on the subtlest true failures hurt more than random flips of equal volume, but the agent stays well above frozen (0.896/0.863).

Flip rate	Difficulty-corr. AUROC / AUPRC	Asymmetric-random AUROC / AUPRC	Symmetric-uniform AUROC / AUPRC
0%	0.948/0.931	0.948/0.931	0.948/0.931
5%	0.947/0.931	0.948/0.931	0.938/0.919
10%	0.946/0.927	0.948/0.931	0.931/0.911
20%	0.938/0.913	0.947/0.928	0.917/0.892

flip probability is a calibrated logistic in the within-failure percentile of s. Notably, predictive entropy is *lower* on true misses, so an entropy-based difficulty would target boundary cases, not the catastrophic misses – hence the Dice-proximity/low-signal proxy. Scoring labels stay clean; target-rate 0 reproduces 0.9478/0.9311 exactly.

Table S25 confirms the reviewer's concern under the correct control: holding the flip structure fixed (asymmetric, failure-only), concentrating flips on the hardest failures is materially more damaging than spreading them at random (-0.016 AUROC, -0.042 AUPRC at 20% vs. asymmetric-random) – difficulty correlation, not flip volume, drives the harm. It does not, however, collapse the agent: even at 20% it stays $+0.052$ AUROC / $+0.133$ AUPRC above the frozen no-memory baseline, retaining $\sim$76% of the clean AUROC gain.

The agent on the unfiltered stream. All headline metrics are computed over tumor-relevant slices, a subset defined by ground truth (Sec. 1 of the main paper). A deployed agent instead ranks the *unfiltered* stream, so we score the same fusion agent and the same frozen baseline on all slices of the external-Africa cohort, seed-averaged over the same ten calibration seeds (Table S26). Two things follow. First, the filtered rows reproduce Table 1 of the main paper exactly, so the two streams are scored by an identical pipeline. Second, the agent's lead over frozen is *larger* unfiltered ($+0.115$ AUROC / $+0.150$ AUPRC) than filtered ($+0.052$ / $+0.069$), while both methods degrade in absolute terms as failure prevalence falls from 0.398 to 0.203 over roughly twice as many slices. The main paper's claim – that the agent still leads frozen on the harder operating point, but both degrade as prevalence falls – therefore holds, and the filtering choice is conservative with

respect to the *gain* we report. These are AUROC and AUPRC only; we did not re-run the selective-review budget metrics on the unfiltered stream.

Table S26. Filtered vs. unfiltered external-Africa stream (spiking segmenter, calibrate PEDs+GLI cap-5, ten seeds, mean±std). The filtered rows reproduce main-paper Table 1. The agent's margin over frozen is wider on the unfiltered stream, though both degrade as prevalence falls.

Stream	Method	AUROC	AUPRC	Prevalence	Rows/seed
Filtered (tumor-relevant)	Fully-online fusion (agent)	0.9478 ± 0.0042	0.9311 ± 0.0043	0.398	6,459
Filtered (tumor-relevant)	No-memory, frozen	0.8961 ± 0.0351	0.8625 ± 0.0434	0.398	6,459
Unfiltered (all slices)	Fully-online fusion (agent)	0.8734 ± 0.0043	0.6543 ± 0.0109	0.203	12,628
Unfiltered (all slices)	No-memory, frozen	0.7585 ± 0.0421	0.5045 ± 0.0818	0.203	12,628

20 An Empirical Trend: the Memory Gain is Predictable from Controller Strength

The class-balanced memory adds little over an adapting controller on most external settings but resolves a large gain on nnU-Net. Rather than treat this as an unexplained exception, we characterise *when* the memory resolves. We assemble 60 settings (segmenter × cohort-config × review-budget cap) spanning the spiking segmenter, nnU-Net, SwinUNETR, and a true Swin deep-ensemble, external and broad configurations, caps 5/10/20. For each we run the full online protocol and record x = the online no-memory controller's AUPRC and y = the memory-fusion increment over it, seed-averaged over ten seeds. Anchors reproduce exactly (spiking external-Africa 0.874→0.919; nnU-Net broad cap-20 0.461→0.643).

The memory increment scales with controller *weakness*: across all 60 settings Pearson $r = -0.68$ (Spearman -0.70) between controller AUPRC and ΔAUPRC, and this is not an nnU-Net effect – *excluding* nnU-Net the correlation *strengthens* to $r = -0.95$ ($n = 45$), with the spiking segmenter, SwinUNETR, and the deep-ensemble each independently tracing the same monotone curve across their cohorts (Table S27). In-distribution the trend is nearly linear (broad-pooled, $n = 12$: $r = -0.98$, $y \approx 0.162 - 0.164\,x$), and an in-sample threshold "expect a resolvable gain (ΔAUPRC>0.05) iff controller AUPRC <0.463" separates settings at 88% (this cut is fit on the same 60 settings and is not out-of-sample validated; the linear fit above crosses ΔAUPRC=0.05 at $x \approx 0.68$, so treat 0.463 as a discriminative threshold, not the trend's zero-crossing). **Honest caveat:** nnU-Net *external*-Africa inverts the rule – the controller is weak (0.46–0.55) yet memory *hurts* (-0.01 to -0.03) – because the class-balanced memory itself does not transfer under that domain shift. The precise statement is therefore: *the memory helps in proportion to controller weakness when the memory is in-distribution*; external transfer is where it can break.

Table S27. The memory increment scales with controller weakness across segmenters (representative rows; full 60-setting sweep in the code release). x = online no-memory controller AUPRC; Δ = memory-fusion increment over it. Low-x (weak controller) settings gain; high-x settings do not; the nnU-Net external row is the transfer-failure exception.

Segmenter	Config (cap)	Ctrl AUPRC (x)	ΔAUPRC (y)
SwinUNETR	broad@PEDs (20)	0.903	+0.017
Spiking	broad-pooled (20)	0.934	+0.004
SwinDeepEns	broad@PEDs (20)	0.942	+0.011
SwinDeepEns	broad@GLI (20)	0.712	+0.053
nnU-Net	broad@PEDs (20)	0.703	+0.103
SwinUNETR	external-Africa (10)	0.404	+0.096
SwinDeepEns	broad@Africa (20)	0.415	+0.093
nnU-Net	broad@GLI (20)	0.370	+0.189
nnU-Net	broad-pooled (20)	0.461	+0.182
nnU-Net	external-Africa (20)	0.545	−0.026 *(transfer fails)*

A prevalence caveat on this trend. Both axes are AUPRC, which is prevalence-dependent, and base failure rates across the 60 settings vary widely (from 0.045 to 0.836 across our cohort×segmenter cells; Table S23). Part of the x–y correlation may therefore be driven by prevalence rather than by controller strength alone: settings with rare failures depress both the controller's AUPRC and the headroom above it. Two observations argue the trend is not purely a prevalence artifact – each segmenter independently traces the same monotone curve *across its own cohorts*, and excluding nnU-Net the correlation strengthens to $r = -0.95$ – but we have not partialled prevalence out, so we report this as an empirical trend rather than a prevalence-adjusted law.