Giving experts a way to say "not mine": ten pre-registered selector and abstention experiments

W. Schulz

Nov 9, 2026

Abstract

Frozen per-task experts on Split-MNIST vote confidently for tasks they never saw. Ten experiments, frozen before running and run once over 5 seeds, asked what stops it. A selector that never reads expert confidence (RanPAC over streaming-PCA features) reached 0.926 class-IL accuracy at ε 0.067. A wrong-answer penalty failed inside the expert and worked at the selector.

Written 2026-11-09, before any run. Follows `EXPERIMENTS-selector-epsilon.md` (SE1–SE7). The literature review `NeuroGenesys/reports/Stopping confident wrong expert votes.md` produced a prioritized 10-experiment plan; this file freezes its predictions and kill criteria. Each experiment is run once, over 5 seeds, and reported in full whether it holds or dies.

Question. Frozen per-task experts on Split-MNIST vote confidently on inputs from tasks they never saw (selector ε≈0.39 raw pixels, 0.28 on streaming-PCA features). Which of these stops it: (i) a selector that never reads expert confidence, (ii) an abstain option at the selector, (iii) calibrating experts onto one scale, (iv) training experts against other-task inputs, (v) a penalty for wrong answers without other-task inputs, (vi) self-made negatives, (vii) reconstruction gating, (viii) deciding the task over a batch?

Harness. `cl/` numpy MLPs, `experiments/selector_abstain/`. Split-MNIST, 5 tasks × 2 digits, 500 train / 100 calibration / 300 test images per digit, drawn disjointly from the MNIST files. K=5 bootstrap experts per task (hidden 40 on raw pixels, 32 on PCA features, 10 epochs, lr 0.1). Streaming PCA r=100 fit on the full unlabelled training stream (as SE5). Seeds 0–4 (data draw and inits). The calibration split is used only where an experiment says so; anything fit on it is declared as using stored labelled data. "Buffer" means m stored training images per task.

Metrics (every experiment, mean ± sd over seeds): ε = fraction of test inputs where some wrong task's score beats the correct task's; A_sel = task-selection accuracy; p_within = oracle-task class accuracy; ACC = class-IL accuracy; task-OOD AUROC where defined.

Experiments, predictions, kills

#	Script	Prediction	Kill
0	`sa00_calibration_diagnostics.py`	Experts are calibrated on their own task (ECE < 0.05) but overconfident on mixed inputs: the $\geq$0.95 confidence bins hold far fewer than 50% own-task inputs. A shared temperature leaves ε exactly unchanged; a per-expert isotonic map from score to P(mine) (fit on the calibration split) lowers ε.	Mixed-input reliability is good (ECE < 0.05) and task AUROC $\geq$ 0.95
1	`sa01_statistical_selectors.py`	On PCA features, shrunk per-class Mahalanobis (FeCAM-style, γ=0.5) and RanPAC (random ReLU projection M=2000 + ridge, λ picked on calibration) beat global NCM (H8, 0.871) by 2–6 ACC points, with lower ε and a flatter $\varepsilon(T)$, T=2..5.	No variant beats NCM by $\geq$ 1 ACC point averaged over seeds
2	`sa02_abstention_threshold.py`	Risk–coverage: abstaining on the MSP selector removes few wrong selections (they are confident); a density selector's selective risk falls much faster. Kalai score (+1 / 0 / $-t/(1-t)$, $t \in \{0, .5, .75, .9\}$) higher for the density selector at every $t > 0$.	MSP area under risk–coverage $\leq 1.1 \times$ the density selector's

#	Script	Prediction	Kill
3	`sa03_scale_calibration.py`	(a) Per-expert affine sigmoid(α s + β) fit on a 20- or 200-image-per-task buffer raises ACC by a few points over raw MSP / max-logit / energy. (b) Buffer-free percentile-rank of each expert's score against its own in-task calibration scores gives a smaller gain.	(b) gains < 1 point while (a) gains $> 2 \rightarrow$ recorded as "buffer-free calibration does not substitute" (finding, not a method kill)
4	`sa04_not_mine_head.py`	A third "other" output trained on a buffer of other tasks' digits (m = 10/50/100 per task; past tasks only during training, then ROW-style head re-tune on the final buffer) cuts ε sharply (Henning 100/task: ~9% wrong-task). A post-hoc "mine vs other" logistic head on frozen expert features does worse (feature collapse).	ε not below the best #1 selector's ε at any m
5	`sa05_penalty_only.py`	Experts trained with a 2-class + abstain output under the Deep-Gamblers loss (payoff o = 1/t, t $\in$ {.5, .75, .9}) on own-task data only: ε essentially unchanged vs plain experts; in-task coverage shrinks.	ε falls by $> 30\%$ relative with no other-task data

#	Script	Prediction	Kill
6	`sa06_self_negatives.py`	CSI-style "shifted" class (rot 90/180/270, h-flip of own-task images) recovers much of #4's gain with no stored data. Generic outlier exposure (pixel-permuted and noise images, OE loss to uniform) helps less.	CSI arm gains < 3 ACC points over plain experts
7	`sa07_reconstruction_selector.py`	Per-task PCA subspace (k=20) reconstruction error, normalized by each task's own calibration error, beats MSP but trails the best #1 selector.	Below the MSP selector
8	`sa08_batch_inference.py`	Summing selector log-scores over same-task batches drives ε near 0 by batch 8–16 for the density selector; MSP converges more slowly.	Density selector ε at batch $32 > 0.05$
9	`sa09_confusion_structure.py`	The per-pair wrong-win matrix is highly non-uniform; an independence model with one homogeneous pairwise rate over-predicts ε for larger task subsets, while the heterogeneous model ($1 - \Pi(1 - p_ij)$) fits better. Also run on synthetic class-IL with T=10 (SE6 setup).	Off-diagonal coefficient of variation < 0.25 (near uniform)

Declared choices made before running: FeCAM-style shrinkage γ=0.5 without Tukey transform (PCA features are signed); relative Mahalanobis subtracts a constant per input, so it cannot change an argmax selection and is evaluated only for abstention (#2) and task AUROC; RanPAC λ chosen on the calibration split from {0.01, 0.1, 1, 10, 100}; isotonic and logistic maps implemented in numpy. The MNIST pairing (0/1, 2/3, ...) caps T at 5, so ε(T) on MNIST uses the first T tasks.

Results (run once, 2026-11-09, seeds 0–4, mean ± sd)

Logs and JSON: `results/selector_abstain/`. Re-run: `experiments/selector_abstain/run_all.sh`. Baselines in this harness: MSP consensus selector ACC 0.633 ± 0.009 (ε 0.367) on raw pixels and 0.786 ± 0.004 (ε 0.213) on PCA features, matching SE5/H4 ACC (0.635 / 0.788). ε here is scored on the K-expert consensus; SE1–SE7's 0.28 PCA figure was the single best expert.

#	Verdict	Result
0	Held (raw); ≥0.95-bin share prediction failed on PCA	Own-task ECE 0.011 (raw) / 0.023 (PCA); mixed-input ECE 0.697 / 0.646; task AUROC 0.822 / 0.908. Of raw inputs scored ≥ 0.95, only 34% were the expert's own task (PCA 69%, above the predicted < 50%). Shared temperature: ε 0.367 → 0.368. Per-expert isotonic P(mine): ε 0.367 → 0.336, ACC +3.0 (raw); PCA +0.5.
1	Held	ACC / ε: NCM 0.857 / 0.131, FeCAM-style 0.892 / 0.095, RanPAC 0.926 / 0.067 (best − NCM = +6.9, at the top of the predicted range). ε(T=2..5): MSP 0.071→0.213, NCM 0.039→0.131, RanPAC 0.016→0.067. Hybrid (task by selector, class by experts) within 0.006 of direct.
2	Held	AURC: MSP 0.098 (margin) vs density 0.025 (FeCAM margin), 4× lower. Risk at 70% coverage 0.123 vs 0.023. Kalai score at t = 0.75: MSP +0.354 vs density +0.680; at t = 0.9: +0.114 vs +0.545 (71% coverage). Margin beat top score and relative Mahalanobis for the density selector.
3	Finding fired: buffer-free calibration does not substitute	Affine sigmoid on a 200/task buffer: raw max-logit 0.619 → 0.673, energy 0.618 → 0.673 (+5.4); raw MSP got *worse* (0.633 → 0.585), because MSP saturates near 1. Buffer-free percentile rank lowered ACC everywhere (−1 to −5 points). PCA gains ≤ 2 points.

#	Verdict	Result
4	Kill fired	"Other" output with past-task buffer + head re-tune, raw: m = 10 / 50 / 100 → ACC 0.733 / 0.847 / 0.871, ε 0.264 / 0.149 / 0.124 (Henning predicted ~0.09). Large gain over plain experts, but ε never fell below RanPAC's 0.067 (or FeCAM's 0.095) with no buffer, so at these sizes training-side abstention does not beat a good selector. Post-hoc head on frozen features worse, as predicted (raw m = 100: ε 0.303; m = 10 worse than plain). PCA m = 10 hurt (ε 0.272).
5	Held (penalty alone does not help; at high t it hurts)	Deep Gamblers, own-task data only. t = 0.5: ε 0.367 → 0.359 (−2%), abstain never fires. t = 0.75: ε 0.564 (+54%), abstains on 18% of own-task inputs and 20% of other-task inputs. t = 0.9: ε 0.739, in-task coverage 0.24. The abstain output fires almost equally on own and other tasks, so it carries no competence information.
6	Held raw; kill fired on PCA (+2.98 vs 3.00)	CSI-style shifted class, raw: ACC 0.633 → 0.760 (+12.7), ε 0.234 (MSP readout 0.769). PCA: +3.0 / +4.0 (MSP readout). Generic outlier exposure: +2.3 raw, +1.1 PCA.
7	Held	Per-task PCA reconstruction, normalized as pre-registered: raw ACC 0.866 (ε 0.124) vs MSP 0.633; PCA 0.802. Secondary arm (in the script before running, not the pre-registered primary): unnormalized error, raw 0.915 (ε 0.075), PCA 0.888. Normalizing by own-task median error hurt.

#	Verdict	Result
8	Held	Density selector ε by batch size 1/2/4/8/16/32: 0.095 / 0.044 / 0.011 / 0.002 / 0.000 / 0.000. MSP: 0.213 / 0.133 / 0.052 / 0.012 / 0.002 / 0.000, i.e. slower but also converges, so MSP's errors are sample-level, not task-level.
9	Non-uniform held; heterogeneous-fits-better failed	Off-diagonal CV 0.687 raw, 0.573 PCA (max pair 0.39 vs mean 0.15). Both independence models over-predict ε: full-set measured 0.367 vs homogeneous 0.488 / heterogeneous 0.471 (raw); synthetic T = 10: 0.551 vs 0.878 / 0.871 (synthetic CV 0.20, near uniform). Heterogeneity barely helps (MAE 0.031 vs 0.034). Experts of the same wrong task agree on which inputs they steal (phi 0.77 raw).

What the campaign says

1. Change the selector, not the experts. The best result is a selector that never reads expert confidence: RanPAC over streaming-PCA features, ACC 0.926 ± 0.004 (ε 0.067), above GDumb (0.859) and H8's NCM. It also has the flattest ε(T). Caveat inherited from SE5/H8: the PCA basis is fit on the whole unlabelled stream, and RanPAC's ridge λ is chosen on a labelled calibration split.
2. The penalty has to sit at the selector and act on an independent score. A reject threshold on the density selector's margin keeps 71% coverage at t = 0.9 with Kalai score +0.545; the same threshold on expert confidence gets +0.114 at 21–36% coverage.
3. A penalty without other-task data does nothing, and at strict thresholds it hurts (SA5): the abstain output fires uniformly on own and other tasks. With other-task data (SA4) the same idea works (0.633 → 0.871), and self-made negatives recover about half of that with none (SA6, raw).
4. Calibration is fine where it was trained and badly wrong elsewhere (SA0: ECE 0.011 vs 0.697), and the usual fixes barely help: shared temperature cannot change selection, per-expert isotonic P(mine) gains 3 points, buffer-free percentile calibration loses.
5. Grouped inputs nearly solve selection: 8–16 same-task inputs take ε to ~0 for every selector.
6. ε(T) grows sub-linearly because confusions overlap on the same inputs, not because a few pairs dominate. Independence models over-predict ε at every task count. (The overlap is inferred from that over-prediction; cross-task co-occurrence was not measured directly.)
7. A fully task-local selector works on raw pixels: unnormalized per-task PCA reconstruction (each task uses only its own data, no future stream) reaches 0.915. It was a secondary arm.

Seeds 5; uncorrected multiple comparisons across ~40 arms; Split-MNIST only.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/selector-abstention