AI That Knows When Not to Act: Trustworthy Physical AI at the Edge

A deployed model can stay confident after the product, lighting or process changes. Capsa adds a risk signal alongside the prediction, so an edge system can act on familiar inputs and route the rest to human review.

Running live on the i.MX 8M Plus NPU full INT8 · TensorFlow Lite 2.19 · eIQ VX delegate

The result in 30 seconds

44 / 151

known-good images auto-accepted

`i.MX 8M Plus`
`full INT8 · TFLite 2.19`

145 / 150

defect images routed to a human

`i.MX 8M Plus`
`full INT8 · TFLite 2.19`

+0.272 ms

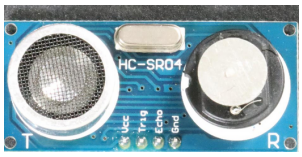
added NPU inference time (+3.26%)

`i.MX 8M Plus`
`p50 of 200 runs`

Public VisA PCB1—3 test data, not a production line. Five of 150 defect images were accepted by the policy. Customer clearance and escape rates will depend on prevalence, product mix, process changes, deployment backend and the selected tolerance. These are not yield, verification-labour, production-PPM or safety figures.

What the gate sees

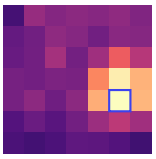
Defective board — routed to a human



Input

Capsa risk map

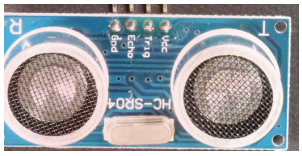
Defect mask (scoring only)



Native 7×7 risk

Known-good board — accepted without review

Input



Capsa risk map

Defect mask (none — good board)

Native 7×7 risk

The classifier is trained on known-good images and has no defect head; masks score the benchmark, never the gate. The mask panel keeps the photograph only inside the ground-truth defect — the good board's is black because there is nothing to mark. On the risk maps opacity carries risk: clear where the gate is comfortable, red where it is not. Both grids (right) share one scale, brighter is higher risk — **seven** of the 49 cells cross the policy's own hot-cell threshold on the defective board and **none** on the known-good one, and that count over 49 is the abstain score the gate acts on.

VisA PCB1 (CC BY 4.0), same crop and same scales, both anchored to the policy's own per-cell risk threshold — neither stretched to its own image's range, which is what would make every board look like it has a hot spot. The deployed INT8 graph, which is bit-exact on the board's CPU. Illustrative of the deployed gate; not measured on the NPU.

Cost, quality and how an evaluation runs

The operating point, and the grid behind it

Calibrated under a **declared 10% escape tolerance** — a one-sided 95% upper bound, not a predicted outcome. The bound behind this point is 9.4%; the escape realized on the held-out half was 3.3%. The headline figures are point estimates for the i.MX 8M Plus on the full INT8 graph. Its two execution paths are not distinguishable on this data: the defect result is identical on both (5 / 150, 3.3%, 95% CI 1.1—7.6%), and the acceptance intervals overlap, each point estimate falling inside the other's. Quantization and precision shift score magnitude between them, which is why a threshold is calibrated on the path that will run it — the per-backend rows are below.

Backend	Declared	Bound	Cleared		Realized escape		Held unnec.
Host INT8	5%	4.1%	35 / 151	23.2%	4 / 150	2.7%	4 / 151
Host INT8	**10%**	**9.4%**	**44 / 151**	**29.1%**	**5 / 150**	**3.3%**	**4 / 151**
Host INT8	20%	19.5%	65 / 151	43.0%	16 / 150	10.7%	4 / 151
Board NPU	5%	4.1%	29 / 151	19.2%	3 / 150	2.0%	1 / 151
Board NPU	10%	9.4%	37 / 151	24.5%	5 / 150	3.3%	1 / 151
Board NPU	20%	19.5%	66 / 151	43.7%	15 / 150	10.0%	1 / 151
Host FP32	5%	4.1%	29 / 151	19.2%	1 / 150	0.7%	2 / 151
Host FP32	10%	9.4%	55 / 151	36.4%	11 / 150	7.3%	2 / 151
Host FP32	20%	19.5%	79 / 151	52.3%	17 / 150	11.3%	2 / 151

All three backends, all three declared tolerances, one locked pass. **Only the bold row carries the predeclared comparison**; every other cell is descriptive. Host FP32 at a declared 10% overshoots to 7.3% realized escape — which is why a threshold cannot be ported between backends. "Held unnec." is known-good boards sent to a person needlessly.

At the primary point: accept-rate 95% CI 22.0—37.1% host, 17.9—32.2% board; escape 95% CI 1.1—7.6% on both (Clopper—Pearson). No tolerance tighter than 5% is declarable here — with zero escapes in 150 calibration defects the 95% bound is still 1.98%, and a 1% bound needs at least 299.

Illustrative projection, not a measurement. Applying the host-INT8 accept rate to a hypothetical stream that is 1% defective and 0.2% unseen-product gives **28.8%** of inputs auto-accepted. That is arithmetic under an assumed prevalence — not a measured clearance rate, not a labour saving, and not the share of this test set that was cleared.

What it costs, and what it leaves alone

Measurement	Unwrapped	Capsa	Difference
i.MX 8M Plus NPU inference (p50, 200 runs)	8.34 ms	8.61 ms	+0.272 ms, +3.26%
Deployed artifact size	2.71 MB	3.15 MB	+0.447 MB, +16.5%

- Same full-INT8 TensorFlow Lite / eIQ VX deployment path.
- No cloud calls in the inference path.
- Host-TFLite INT8 and board-CPU tensor values were bit-identical for the deployed artifact.
- Everything numerical in this section was measured on i.MX 8M Plus.
- Policy behavior was regression-tested against the reported decision logic.

The same TensorFlow Lite export format is consumed by eIQ Neutron on i.MX 93 and i.MX 95. **This brief does not claim that the model or performance numbers have been validated on those devices.** Equivalence of the Capsa export path and a hand-built export has been shown for this VisA model, on this backend; it is not claimed in general.

Localization and the detector reading

Localization quality by execution context

Execution	Pixel-AUROC	95% CI
FP32 host	0.8874	0.849—0.916
INT8 host	0.8751	0.832—0.906
INT8 board VX / NPU (deployed)	**0.8714**	0.827—0.904

Same 150-image test half. Intervals are 2,000-resample image-clustered bootstraps; the image, not the pixel, is the unit of resampling. The three rows are three backends, not three runs of one.

Risk ranking survived NPU deployment: Spearman 0.992 between host INT8 and board VX. Quantization rescales absolute risk magnitude but leaves image ordering essentially intact.

Quality-only comparison. The board-VX Capsa map reached 0.871 pixel-AUROC versus 0.793 for host-FP32 PaDiM. PaDiM requires a separate feature pipeline that the deployed graph does not emit. PaDiM was host FP32, was not deployed, and was not latency-measured — this compares localization quality only, and is not a triage, throughput or review-reduction comparison.

The same gate, read as a detector

The sections above declare an escape tolerance and report how much review is skipped. The same threshold family reads from the other end: declare a review budget, report what fraction of defects is caught. Both come from the same locked pass; this reading is descriptive and carries no comparison.

Declared budget	Defects caught		$\geq$ at 95% conf.	Stream reviewed
1%	unreachable at any threshold			
2%	29 / 150	19.3%	14.2%	2.8%
5%	46 / 150	30.7%	24.5%	8.2%
10%	67 / 150	44.7%	37.8%	20.1%
20%	88 / 150	58.7%	51.6%	29.4%

Host INT8, on a stream where 1 board in 100 is defective. Read the confidence bound rather than the point estimate.

A negative result. Every reachable cell overshot its declared budget on the held-out half. The abstain action flags images whatever their verification score, putting a 2.8% floor under the budget, and the false-abstain rate moves from 1.3% on calibration to 2.6% on test. A review budget of a few percent cannot be calibrated from 150 images, whatever the score. That is a limit of the sample rather than of the method.

The two framings are one operating point read from opposite ends. At the primary point the gate catches **145 of 150 defects (96.7%)**, and to do it sends 107 of 151 known-good boards (70.9%) to a person. That is the same fact as "clears 29.1% at 3.3% escape".

Protocol, relevance and methodology

From public benchmark to your line

1. **Bring the deployed model and backend.** The model you already ship, and the NXP part it runs on. We wrap the graph you have.
2. **Provide process images, including known-good production.** Capsa is fitted without defect masks. Labelled defects are required to select and validate an operating point with a declared escape tolerance.
3. **Name the change condition and tolerance.** A new product variant, supplier, lighting rig, paste lot or facility — and the escape tolerance you are willing to declare.
4. **Lock the threshold on the deployment backend and evaluate once on held-out data.** You get a coverage-versus-error curve and a proposed human-review policy.

Measured lesson. Quantization preserved ranking but changed score magnitude. Transferring a threshold across backends produced **21.3% escape against a 5% target**, so thresholds must be calibrated and locked on the backend that will actually run them.

Capsa decides when an image needs review. It is not a safety function and does not set a PPM rate.

Relevance across the NXP customer base

Demonstrated — electronics AOI / PCB inspection. An existing human-verification workflow, public PCB data, and the i.MX 8M Plus proof above.

Adjacent — automotive parts, battery cells, pharmaceuticals, cosmetics. Same problem shape; each requires evaluation on customer data. The PCB benchmark does not carry over as a result.

Adjacent — wearables and physiological sensing. Separate board-level work on physiological time-series models. No numerical claim is published for this class.

Adjacent — condition monitoring and predictive maintenance. Earlier host-side work, including eIQ Time Series Studio — host-side only, and not a board-deployment claim.

Methodology

Dataset and split. Public VisA PCB1—3 (CC BY 4.0), held-out half. Policy denominators are 151 known-good and 150 defective images. Localization is scored on the 150 defective images only.

Stack. Full INT8, TensorFlow Lite 2.19, eIQ VX delegate, i.MX 8M Plus. Latency is the p50 of 200 runs after 30 warm-ups, against a matched unwrapped export sharing the architecture and INT8 quantization recipe, so the delta isolates the wrapper rather than the build.

Intervals. Pixel-AUROC intervals are 95% image-clustered bootstraps over 2,000 resamples; policy counts are Clopper—Pearson. With 150 defect images, 5 escapes span 1.1—7.6%.

Declared tolerance versus realized rate. A declared tolerance is a one-sided 95% upper bound from the calibration half — a promise about the worst case, not a prediction. The bound is 9.4%; the realized rate was 3.3%.

Thresholds are backend-specific. Capsa risk is an absolute magnitude and quantization rescales it. Thresholds must be locked on the backend that runs them. This is what separates the two acceptance figures. The INT8 graph is bit-exact on the board's CPU, so 44 / 151 is a board result. The VX delegate preserves the ranking (Spearman 0.992) and routes the same 5 / 150 defects, but it rescales the magnitude, so the same declared tolerance puts its threshold at a different quantile and it clears 37 / 151. The gap is where the threshold lands rather than a loss of signal — a precision and scaling effect, and the reason to calibrate on the delegate instead of porting a number to it.

Limitations

- **Benchmark, not production.** VisA PCB material is roughly 50% defective; a real AOI station is orders of magnitude cleaner. No false-call labelling, no production data, no external real-PCB dataset. Cross-line generalization is untested.
- **Inherited model-selection bias.** The checkpoint's epoch was chosen by pixel-AUROC on an anomaly pool overlapping the test half (about 0.005 pixel-AUROC), so that experiment is **not confirmatory**.
- **The test half has been examined before.** Thresholds never saw it and the comparison was declared in advance, but the split is not fresh — these results are **locked-but-informed**.
- **The unseen-product result proves nothing about Capsa.** That benchmark is degenerate — the base model's own softmax, and even a mean-brightness rule, separates the unseen product perfectly.
- **Sample size bounds what is declarable.** With 150 calibration defects, the tightest declarable tolerance is about 2%; a 1% escape tolerance is unreachable at this sample size whatever the score.
- **Per-product behaviour varies.** Under one global threshold, per-product escape ranges 0—8%. A pooled bound says nothing about a single product line.
- **PaDiM is host-FP32 only.** Never deployed or latency-measured; the comparison is localization quality only.
- **Human-review policy, not safety certification.** The policy triages whether a human looks at an image — it does not authorise disposal, establish a PPM rate, or certify a safety function. This result is **not a certification**.

Bring us one difficult image and the backend you need to ship.

We will show what the gate sees, then define the data and tolerance for a short evaluation — a scoped piece of work, not a production decision made from one image.

Contact info@themisai.io · themisai.io/nxp-tech-day

Please don't send proprietary process images by email — we'll arrange secure transfer once we're in touch. Every figure here maps to a result file and execution context.