



# The Decision Is One Token

Typed decision models, prompted seats and a labelled head on 829 of our own items

**Paper:** PT-R-2026-040 v1.0

**Author:** PureTensor Ltd

**Date:** 20 October 2026

**Status:** Published

## Abstract

---

Our agents make thousands of small decisions a day. Is this message from the operator a correction? Does this new note supersede an old one? Should this alert page a human? Is this email phishing? Does this document carry a prompt injection? A survey of our own fleet found about 120 such text decision gates, built over the past year from regular expressions, embedding thresholds and full language-model verdicts. In September a new category of model appeared to replace them: small "System One" decision models that take a typed question and return a calibrated answer (a yes/no probability, or a choice with a probability per option) instead of free text. By October there were three open-weight families under permissive licences.

We measured them head to head on our own labelled data: 829 items across seven tasks drawn from real traffic and real gold labels. The arms were two zero-shot decision models (Clef-flash 9B and Strands Decider 2B), an earlier encoder-style decision model (Laya) in two modes, two of our own serving seats used as one-token classifiers with probabilities read from token log-probabilities, and a frozen embedding with a logistic-regression head trained on the labels.

No arm won everywhere. Clef-flash was the best zero-shot model on four of six scored tasks: alert action accuracy 0.747 with every page-worthy alert caught, inbox intent accuracy 0.969 across 12 classes, and injection detection AUROC 0.983 with 88.3% recall at a 5% false-positive rate. Our resident 120B seat, asked for a single letter, was the best zero-shot judge of operator corrections (AUROC 0.925) once we gave it only the operator's turn; adding the previous assistant turn as context lowered it to 0.823. Where labels exist, a logistic head on frozen embeddings matched or beat every zero-shot model on four tasks, and on memory-poisoning payloads it reached 89.4% recall at zero false alarms where the best zero-shot model reached 78.7%.

Clef-flash now serves on one GPU in 19 GB at 38 ms median latency, with bf16 GPU scores matching the CPU evaluation (deepset AUROC 0.982 against 0.983). Its first consumer screens incoming news documents for injection, in shadow. The results are single-run, the gold sets are small (16 positives on the correction task, 8 clean items on the poisoning task), and nothing yet acts on these scores. Section 9 states which claims survive that.



## 1. The Question

A decision gate is any place where software turns text into a branch. Ours range from trivial (a regular expression that flags an operator message containing "no, I said") to expensive (a council of four language-model calls, each 9 to 25 seconds at the median, that decides whether a collected news document is good enough to enter an intelligence brief). In October we inventoried them across the fleet and found about 120: language-model verdicts, regex classifiers and embedding thresholds, written one at a time by whichever agent needed a branch that day.

Most of these gates share three faults. They return text that has to be parsed back into a decision. They give no usable confidence, so there is no principled threshold between "act" and "ask". And the expensive ones run a large model to answer what is, structurally, a one-token question.

The System One category promises to fix all three. A caller sends a state (the text to judge) and a set of typed questions. Each question is a "noui" (a yes/no answer with a probability), a "choice" (one option from a fixed set, with a probability for each), or a "score". The model returns calibrated numbers. The interface began as a hosted commercial service. By 20 October it had become an open-weight category: Cloudflare's Clef 27B and Clef-flash 9B (Apache-2.0, released 30 September, a Qwen3.5 backbone with a typed joint head, served through an SGLang endpoint that claims compatibility with the hosted interface), Amazon's Strands Decider LoRAs from 2B to 26B (Apache-2.0, early October), and Fastino's 486M GLiNER2.5-Decide (Apache-2.0). We did not evaluate the hosted service: it is a new paid dependency, it is hosted outside our infrastructure, and published work documents injection attacks against it.

Our question was narrow. On our own data, with our own labels, does a zero-shot decision model beat what we already have, and where does a cheap trained head beat both?

## 2. The Seven Tasks

Every task came from real traffic with a label we trust, and every arm saw identical inputs. Table 1 lists them.

| Task                          | n   | What is decided                                        | Gold labels                                     |
|-------------------------------|-----|--------------------------------------------------------|-------------------------------------------------|
| Operator correction           | 105 | Is the operator correcting the agent?                  | Human labels from our lesson miner; 16 positive |
| Memory reconcile              | 114 | Relation between two memory notes, 5 classes           | Verdicts of a paid frontier judge               |
| Memory poisoning              | 55  | Does a candidate memory write carry an injection?      | 47 attack payloads, 8 clean writes              |
| Alert action                  | 75  | Page now, notify, auto-heal candidate, or known benign | Hand-labelled alert set                         |
| Inbox intent                  | 64  | One of 12 mail intents, plus phishing and needs-reply  | Hand-labelled mail set                          |
| Injection (deepset)           | 116 | Is this text a prompt injection?                       | Public set: 60 injections, 56 benign            |
| Injection (second public set) | 300 | As above                                               | Public labels                                   |

*Table 1. The seven tasks. Every arm saw identical inputs; the second public injection set is excluded from comparisons because not every arm completed it.*

The reconcile labels deserve a caveat at the outset: they are the verdicts of the frontier model that currently judges note pairs for our memory system, not human labels. A score on that task



measures agreement with the incumbent, not correctness. The alert classes are imbalanced (44 notify, 24 page-now, 5 known-benign, 2 auto-heal). The second public injection set was run in full only by some arms (Clef-flash scored 8 of its 300 items before we stopped it for time), so we leave it out of every comparison below.

### 3. The Arms

We ran seven arms.

Laya, in its default and typed modes, is an earlier encoder-style decision model that we had evaluated in September. Clef-flash 9B and Strands Decider 2B are the two new zero-shot decision models; we gave each the same typed questions and criteria text. All four ran on a 64-core CPU node with no GPU, which is where they would run if they replaced a regex.

Two arms are our own serving seats used as classifiers. On 20 October our resident agent seat served Nemotron-3.5-Super-VL-120B-A12B in NVFP4, and our bulk seat served Nemotron-3.5-Lightning-30B-A3B in NVFP4. For each question we sent the options as lettered lines, asked for exactly one capital letter, set the output limit to one token, and read the probability of each option from the first token's top-20 log-probabilities, renormalised over the option letters. The model is never asked to write a probability. It only has to put mass on a letter.

The last arm needs labels. We embedded each item with Qwen3-Embedding-4B (the embedding model our memory system already serves), trained a class-balanced logistic regression, and scored it with stratified 5-fold cross-validation, so every item is scored by a model that never saw it.

### 4. Results on the Memory Tasks

| Arm                                      | Correction AUROC | Correction recall at 5% FPR | Reconcile actionable agreement | Reconcile actionable AUROC |
|------------------------------------------|------------------|-----------------------------|--------------------------------|----------------------------|
| Resident seat, operator turn only        | 0.925            | 68.8%                       | not run                        | not run                    |
| Resident seat, with prior assistant turn | 0.823            | 50.0%                       | 0.746                          | 0.805                      |
| Clef-flash 9B                            | 0.918            | 56.2%                       | 0.693                          | 0.738                      |
| Strands Decider 2B                       | 0.838            | 25.0%                       | 0.640                          | 0.720                      |
| Laya typed                               | 0.688            | 18.8%                       | 0.544                          | 0.434                      |
| Bulk seat                                | 0.646            | 0.0%                        | 0.518                          | 0.681                      |
| Laya                                     | 0.551            | 6.2%                        | 0.509                          | 0.414                      |
| Embedding + logistic head (labels)       | 0.584            | 0.0%                        | 0.675                          | 0.663                      |

*Table 2. Memory tasks. Recall at 5% FPR is read at the threshold that admits 5% of negatives; actionable agreement collapses five relations into change versus no change.*

"Actionable" on the reconcile task collapses the five relations into the ones that should change memory (duplicate, contradicts, supersedes) against the ones that should not.

The correction task is where the most useful single result appeared. The regex we run today flags 11 of the 16 corrections (recall 0.688) at a precision of 0.44. At the same recall, the resident seat reading only the operator's message reaches a precision of 0.733. The same model with the previous assistant turn prepended scores lower on every measure. We had expected context to help: a correction is a relation between what the agent did and what the



operator said. On this set it did not. Our reading is that the prior turn, often long and full of tool output, dilutes the signal in a one-token answer, and the operator's own phrasing carries most of the evidence. We now ask relational questions on the operator turn alone.

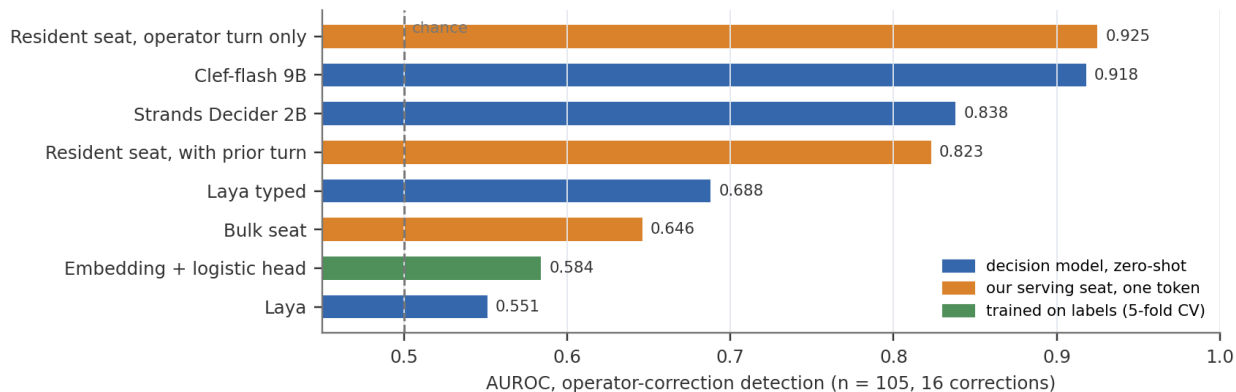


Figure 1. Operator-correction AUROC by arm ( $n = 105$ , 16 corrections). Orange bars are our serving seats read from one token's log-probabilities; context lowered the resident seat from 0.925 to 0.823. Values are in Table 2.

The embedding head fails on this task (0.584). With only 16 positives across five folds, each training fold sees about 13 corrections, and the embedding of a short operator message does not separate "no, use the other host" from "yes, use the other host". Labels help when there are enough of them.

On reconcile, the resident seat agreed best with the incumbent judge (0.746 on whether a pair is actionable). No arm came close to replacing it outright, and Section 9 explains why that matters less than it seems.

## 5. Results on Operations and Mail

| Arm                                | Alert accuracy | Page-now recall | Inbox accuracy (12 classes) | Phishing AUROC |
|------------------------------------|----------------|-----------------|-----------------------------|----------------|
| Clef-flash 9B                      | 0.747          | 1.000           | 0.969                       | 1.000          |
| Resident seat                      | 0.600          | 1.000           | not run                     | not run        |
| Strands Decider 2B                 | 0.387          | 1.000           | 0.891                       | 1.000          |
| Bulk seat                          | 0.200          | 0.292           | 0.844                       | 1.000          |
| Laya typed                         | 0.507          | 0.500           | 0.882 (17 items)            | not scored     |
| Laya                               | 0.080          | 0.000           | 0.734                       | 0.948          |
| Embedding + logistic head (labels) | 0.827          | 0.708           | 0.922                       | not scored     |

Table 3. Operations and mail. Page-now recall is the share of the 24 page-worthy alerts assigned page-now.

Clef-flash is the clear zero-shot leader here. It caught all 24 page-worthy alerts while choosing the right action for three in four alerts overall, and it put 62 of 64 emails in the right one of 12 intent classes. The bulk seat, a 30B mixture-of-experts model that serves most of our high-volume batch work, was close to useless on alert actions (0.200 accuracy, 29% page recall) despite reasonable inbox accuracy. A model that is good at generating text is not, without help, a good classifier of our alerts.

The labelled head shows the trade that recurs through this study. On alerts it had the best accuracy (0.827) but missed 7 of 24 pages (recall 0.708). Zero-shot Clef-flash missed none. For alerts, a missed page costs more than a mislabelled notification, so accuracy is the wrong



headline. Our September evaluation of Laya fine-tuned on 60 alert labels reached 0.907 on the same set; that is the strongest alert result we have, and it needed labels.

## 6. Results on Injection

| Arm                                | Poisoning AUROC | Poisoning recall, no false alarm | Deepset AUROC | Deepset recall at 5% FPR |
|------------------------------------|-----------------|----------------------------------|---------------|--------------------------|
| Embedding + logistic head (labels) | 0.910           | 89.4%                            | 0.961         | 80.0%                    |
| Clef-flash 9B                      | 0.840           | 78.7%                            | 0.983         | 88.3%                    |
| Strands Decider 2B                 | 0.902           | 61.7%                            | 0.914         | 63.3%                    |
| Bulk seat                          | 0.854           | 42.6%                            | 0.918         | 46.7%                    |
| Laya                               | 0.854           | 38.3%                            | 0.897         | 75.0%                    |
| Laya typed                         | 0.840           | 14.9%                            | not run       | not run                  |

Table 4. Injection. The poisoning set has 8 clean items, so its operating point allows no false alarm.

The poisoning set has only 8 clean items, so "5% false-positive rate" there allows zero false alarms; we label the column accordingly.

This table is the reason we no longer compare injection screens by AUROC alone. Strands Decider 2B has a better poisoning AUROC than Clef-flash (0.902 against 0.840), but at an alarm rate an operator would tolerate it catches 61.7% of attacks to Clef-flash's 78.7%. AUROC averages over every threshold, including ones no one would deploy. A screen is used at one operating point, and that is where it should be compared.

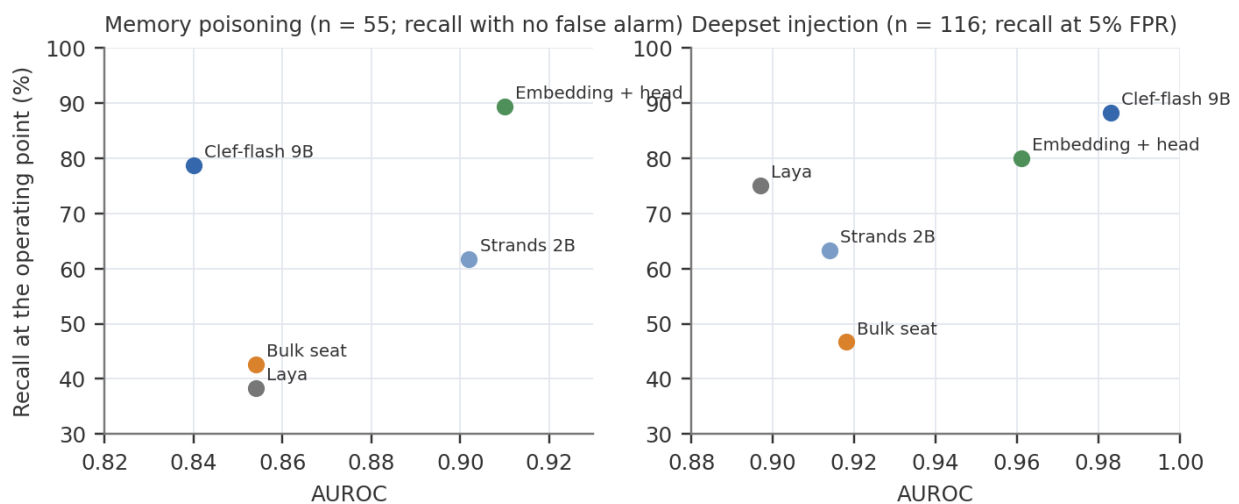


Figure 2. AUROC against recall at the operating point an operator would tolerate. On poisoning, Strands Decider 2B has the second-best AUROC but catches 17 points fewer attacks than Clef-flash at no false alarm. Values are in Table 4.

The other lesson is about calibration. Clef-flash's deepset AUROC of 0.983 sits beside a recall of 10% at the default 0.5 threshold. Its scores rank injections correctly but sit low: the threshold that gives a 5% false-positive rate is 0.026. A decision model that claims calibrated probabilities still needs its threshold set on your own data before it is used.

The labelled head is strongest on poisoning (89.4% at zero false alarms). The poisoning payloads are written against our memory system's own formats, and a head trained on 47 examples of them learns those formats. A zero-shot model has never seen them.



## 7. Cost and Latency

---

On the CPU node, per-item median latency ranged from 0.16 to 1.6 seconds for Laya, 0.66 to 8.9 seconds for Strands Decider 2B and 5.5 to 18.6 seconds for Clef-flash, depending on input length. The two serving seats, already resident on GPUs, answered a one-token question in 107 to 972 ms. The embedding head costs one embedding call plus a dot product.

These latencies made the deployment decision. A 9B model at 5 to 19 seconds per item on CPU is not a gate. On a GPU it is: we moved Clef-flash onto one workstation GPU in bf16, where it allocates 19.06 GB and answers in 38 ms at the median end to end. Before trusting it we re-scored the deepset set on the GPU: AUROC 0.982 against 0.983 on CPU, and the same 88% recall at a 5% false-positive rate. Strands Decider 2B, at about 5 GB, would have fitted without moving anything, and we measured it first for that reason. On injection it was not close at a usable alarm rate (63% against 88% on deepset), so we moved a service to make room for the larger model.

## 8. What Labels Already Know

---

Two of our gates have outcome labels for free: the system records what happened next. We asked whether those outcomes are learnable from data we already store, with no new model calls.

For the news-document council (four model calls per document, failing about 88% of documents), we trained a logistic head on the 384-dimension embeddings the pipeline already stores, plus a one-hot of the collecting source, with a temporal split (first 70% of 21 days to train, last 30% to test, 5,678 documents). It predicted council failure with AUROC 0.784. Skipping the council for the 30% of documents the head was most sure would fail would have discarded 7 of 193 documents the council sent to review and none of the 19 it passed. For our remediation agent, a head predicting whether a model run would resolve a fault reached AUROC 0.735 when grouped by alert name, against 0.40 for a lookup of past success rates. That second result is advisory only: our remediation agent never asks a model whether to act.

Both figures come from the session record of 20 October. The scripts are retained, but the per-document outputs were not, and the 21-day window has moved on, so we report them without re-derivation and give them less weight than Sections 4 to 6.

## 9. What These Numbers Will Not Carry

---

Every number above is a single run on one day. No arm was run twice, so we have no measure of run-to-run variance, though the seat arms used temperature 0 and the decision models are deterministic at inference.

The gold sets are small where it matters most. The correction task has 16 positives; one more caught or missed moves recall by 6 points. The poisoning task has 8 clean items, which is why its "5% false-positive rate" is really zero false alarms, and why its AUROC values carry wide uncertainty. Differences of a few hundredths of AUROC between arms on these two tasks are inside noise. The claims that survive are the large ones: the resident seat on the operator turn beats the regex at equal recall; Clef-flash catches every page and leads zero-shot injection at a fixed alarm rate; AUROC and operating-point rankings disagree on poisoning; the bulk seat is not a usable alert classifier.

The reconcile task measures agreement with the frontier judge we already use, not truth. An arm could disagree with it and be right.



Our serving seats are not the models in them today. On 20 October the resident seat was Nemotron-3.5-Super-VL; it has since been replaced. The "ask the operator turn alone" result is about a 120B model on one task and may not transfer.

Prompts and criteria text were written once and not tuned per arm. A better prompt could move any arm, the seats most of all.

Finally, nothing in this paper acts. Clef-flash scores incoming documents for injection, and our memory system logs System One judgements beside its existing decisions, but both are shadow-only. In a read-only pass over 283 recent news documents, one scored at or above 0.5 (an article about hidden instructions planted for AI assistants, that is, content quoting injections rather than an attack), and the median score was 0.020. That is a sanity check, not a measurement of production precision.

## 10. What We Changed

---

We adopted four rules. Ask relational questions on the operator's turn alone. Read a model's decision from its token probabilities and never ask it to write a probability; the adapter we built in September did the latter. Compare injection screens at a fixed false-positive rate. Set every decision model's threshold on our own labelled data before it gates anything.

Clef-flash serves on one GPU as our default zero-shot decision model, sharing its slot with an on-demand rewriting model that wakes when called. The shadow logs are scheduled for review on 10 November, when we will have the first production precision figures for the injection screen and for the correction and reconcile judgements. The next experiment is the one this study points at: a labelled head per gate wherever outcome labels already exist, with a zero-shot decision model as the fallback where they do not.

## 11. Methods and Artefacts

---

Models: Clef and Clef-flash (Cloudflare, Apache-2.0, September 2026), Strands Decider (Amazon, Apache-2.0, October 2026), Laya, Nemotron-3.5-Super-VL-120B-A12B and Nemotron-3.5-Lightning-30B-A3B (NVIDIA, NVFP4), Qwen3-Embedding-4B (Qwen Team). Public data: the deepset prompt-injection dataset. Metrics: AUROC by the rank-sum (Mann-Whitney) statistic; expected calibration error over ten bins; recall at a fixed false-positive rate read from the sorted negative scores. The embedding baseline is a class-balanced logistic regression (scikit-learn) scored by stratified 5-fold cross-validation. Seat probabilities are the first token's top-20 log-probabilities renormalised over the option letters, at temperature 0 with a one-token limit. Software: PyTorch, Hugging Face Transformers, SGLang, vLLM, scikit-learn, matplotlib.

The inputs, gold labels, every arm's per-item outputs (including the Strands Decider run and the no-context resident run), the scoring script and the script that re-derived the operating-point figures for this paper are retained.

*PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-040. The measurements were taken on 20 October 2026; per-item outputs for every arm in Sections 4 to 6, the inputs, gold labels and scoring code are retained.*