Can an open-weight decision model match frontier performance for an enterprise?

The measured uplift from task-specific Clef training

Mark J Sykes, AI One Labs | 6 November 2026

Executive summary

Yes, on a well-defined enterprise task, a fine-tuned open-weight decision model can deliver performance in the frontier range while keeping business context under local control. AI One Labs trained Clef on 4,998 synthetic examples and evaluated the selected checkpoint against matched stock Clef and the earlier frontier comparisons.

STOCK CLEF NF4	FINE-TUNED CLEF NF4	GPT-6 LUNA / HIGH
92.55%	96.56%	96.10%
1,827 / 1,974	1,906 / 1,974	Responses API

Historical benchmark: 1,974 examples, 50 enterprise workflow categories. Clef uses the matched native eager evaluator; Luna uses high reasoning.

Training reduced stock Clef's errors by **53.7%**, correcting 86 mistakes while introducing seven. On a separate sealed 999-example test, accuracy rose from **98.50% to 99.90%**. The small historical lead over Luna establishes neither superiority nor formal equivalence; it supports frontier-range observed performance on this task. Luna remained more accurate among jointly valid outputs.

Primary data production cost **$156.93**. Training took **4.07 hours** on one RTX 4090, with approximately **$2.80** of allocated depreciation and assumed electricity. Engineering and serving operations are additional costs. The enterprise case combines task-specific quality, a reliable decision interface and context sovereignty.

Results first: pp. 2–3. Study and training: pp. 4–6. Detailed evaluation: pp. 7–12. Economics and implications: pp. 13–16. Methods, class results and sources: pp. 17–20. Dataset and benchmark construction are covered in the separate Paper 3.

Clef joins the frontier performance range

Fine-tuned Clef correctly classified 1,906 of the original 1,974 enterprise requests. The strongest hosted configuration tested, GPT-6 Luna Responses with high reasoning, classified 1,897 correctly. The difference between them is small; the substantial change is the improvement over matched stock Clef.

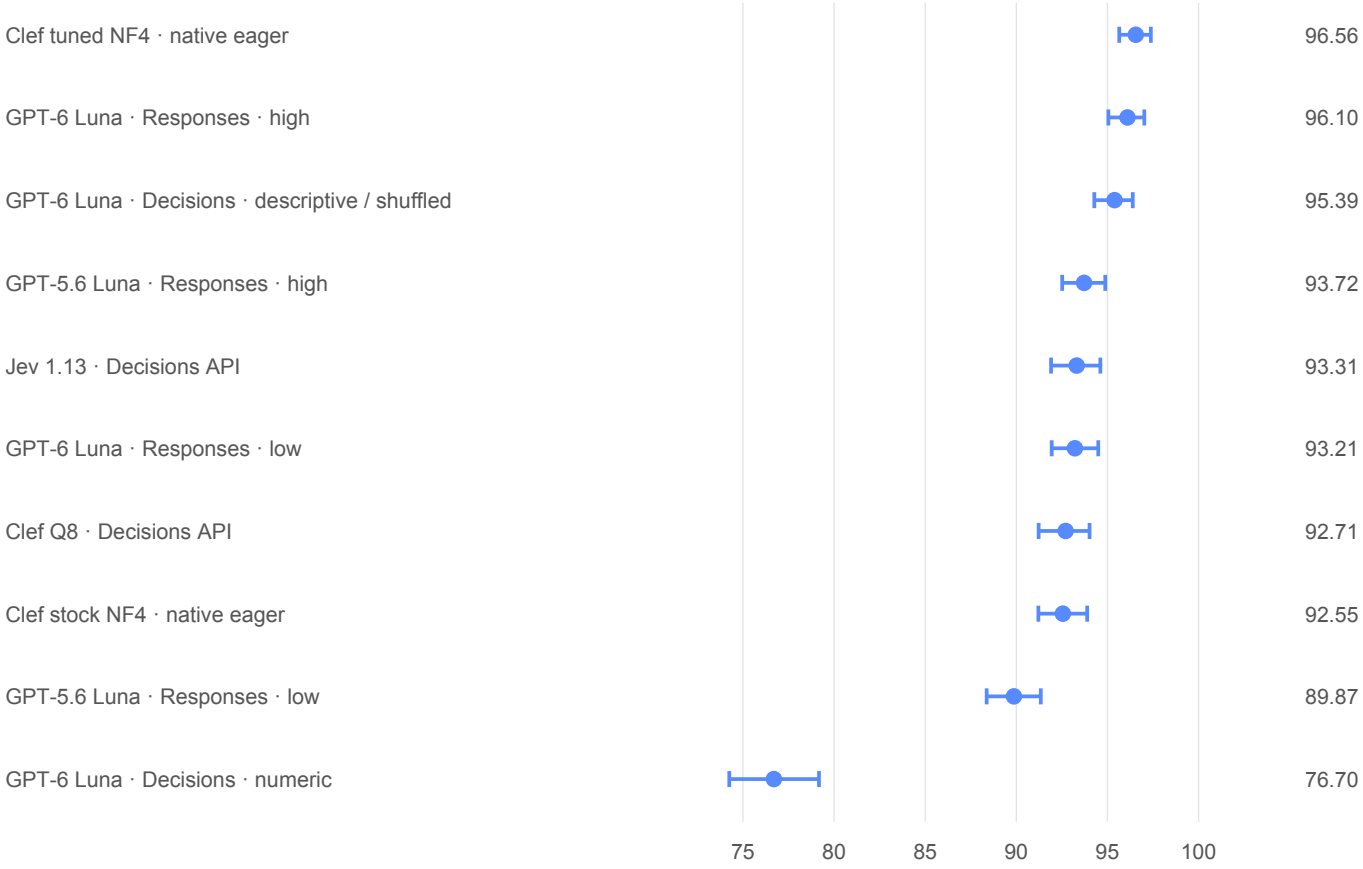


Figure 1. Historical full-cohort accuracy (%), with 95% scenario-component bootstrap intervals. Matched stock and tuned intervals are recomputed with the same method; paired comparisons follow. Axis begins at 72%.

The matched stock score was 92.55%; training raised it to 96.56%. Its 95% paired improvement interval is +2.99 to +5.04 percentage points. Against Luna high, the corresponding interval is −0.36 to +1.29 points. The first comparison supports a clear training gain. The second supports a practical frontier-range claim, without establishing superiority or formal equivalence.

Interface matters. The two Luna Decisions rows differ in both label description and frozen ordering. Their gap is not an isolated shuffle effect. Native Decisions had no settable reasoning level; low and high identify Responses configurations only. Every row is scored on the full cohort, with invalid outputs counted as unsuccessful decisions.

Historical model results retain their recorded configurations. Stock Q8 is the first paper's serving result; matched stock and tuned NF4 use a common native eager evaluator. No shuffle is applied except where named. Sources R1–R3.

The gain is 86 corrections against 7 regressions

The most useful measure of specialisation is what happened to the same requests. Fine-tuning corrected 86 stock errors and introduced seven. That is 79 additional correct decisions, with errors falling from 147 to 68: a **53.7% reduction**.



Same 1,974 examples	Stock correct	Stock wrong
Tuned correct	1,820	86 corrections
Tuned wrong	7 regressions	61 both wrong

Figure 2. The paired result behind the headline: 86 corrections and seven regressions, for 79 additional correct decisions. Both models use matched NF4 native evaluation.

What this changes for an enterprise

A fixed-task classifier can improve substantially without becoming a general-purpose frontier model. Here, a 9B open-weight model learned to distinguish a defined set of enterprise workflows more reliably. Its deployment can remain local, with the same bounded native decision interface.

The remaining errors also matter. Sixty-one requests were wrong under both matched models. Eleven of those 61 cases changed from one wrong category to another, so label changes alone would overstate the improvement. Paired correctness makes the benefit and the regressions explicit.

The separate sealed test supports the same direction of effect: stock made fifteen errors and the selected model made one. That test comes from the training recipe’s distribution; the original benchmark measures transfer into a different, previously used cohort. Together they provide stronger evidence than either result alone.

Historical paired interval: +2.99 to +5.04 percentage points, resampling complete scenario components. Results are conditional on the fixed reference labels. The original benchmark is a development-used regression asset, not a newly sealed test. R2–R3.

One checkpoint, two final comparisons

The task is to identify the current primary workflow from fifty ContextOne sample categories, using a request and its visible context. Models classify the action being requested; they do not execute it. The category meanings and decision policy are fixed across the experiment.

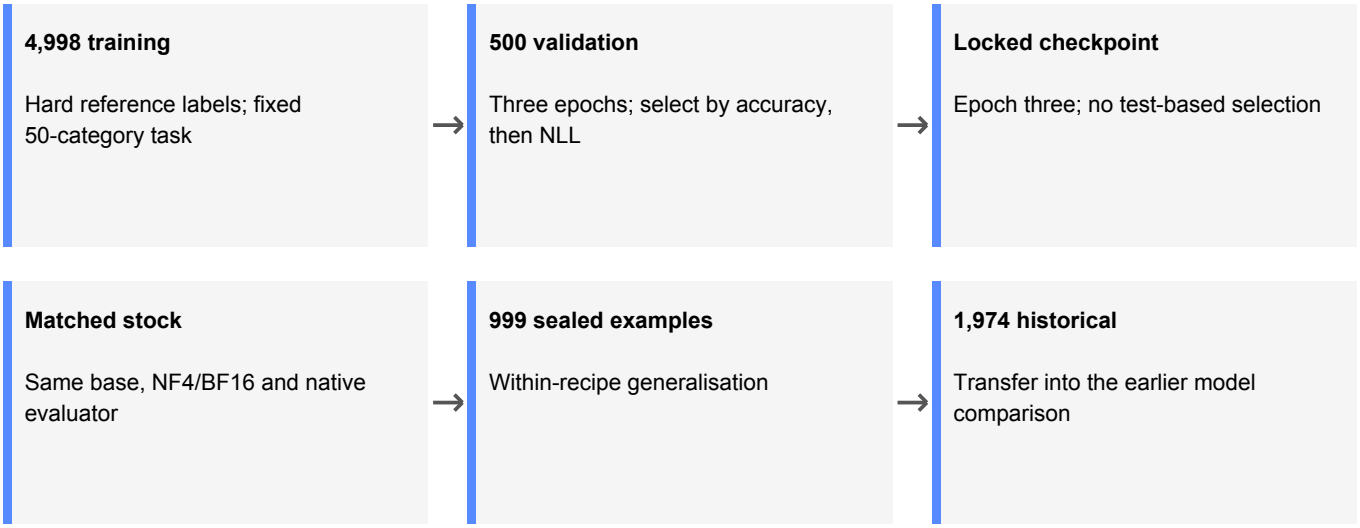


Figure 3. The primary study. The selected checkpoint is evaluated on both test cohorts; the two tests are separate branches, not stages of checkpoint selection.

Dataset	Rows	Purpose
Synthetic training	4,998	Hard-label updates to Clef; all 50 classes represented.
Validation	500	Chooses among three epochs by accuracy, then lower NLL, then earlier epoch.
Sealed synthetic test	999	Final stock-versus-tuned comparison within the generation distribution.
Historical benchmark	1,974	Reuses the exact public inputs and provisional references from the earlier multi-model comparison.

Only validation selects the checkpoint. The historical cohort contains 496 three-variant scenario families plus 486 separate boundary cases, so its uncertainty estimates keep whole families together. It had already been inspected during prior development. The 999-example split supplies a separate sealed test under the same synthetic recipe as training.

Training labels were author-assigned and subject to automated review and curation, not individual human adjudication. The historical labels are also provisional synthetic references. Paper 3 explains their construction, review and separation in detail. This paper evaluates the frozen packs and reports what the resulting model achieved.

No historical benchmark example or related source-history row was eligible for the new training corpus. Taxonomy, membership, encoder and file identities were frozen. R1–R3; companion Paper 3.

Preserve the native decision architecture

Clef was trained as a native classifier. Its original decision head was retained and updated jointly with low-rank language adapters. The student received public prompt/context, the fixed policy, all fifty choices and one hard label. Author explanations and teacher probabilities were not training inputs.

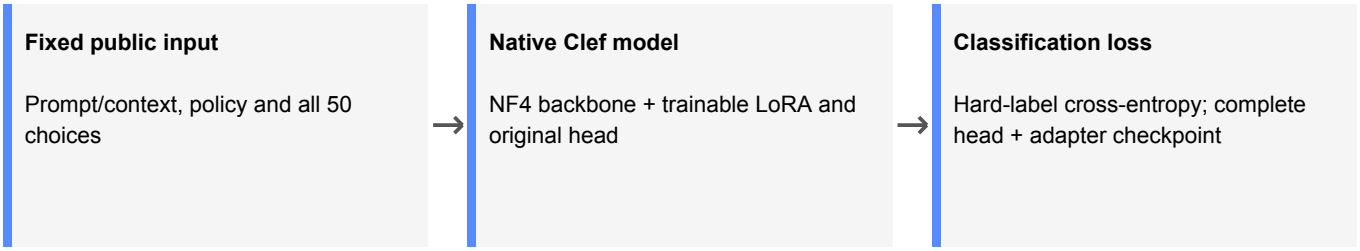


Figure 4. The pretrained decision head is retained and trained jointly with low-rank adapters. This is native classification training, not label-string generation.

Setting	Executed value
Base	Cloudflare/clef-flash, fixed revision, approximately 9B parameters.
Arithmetic	Double-quantised 4-bit NF4 backbone; BF16 forward compute; FP32 trainable head/adapter tensors.
Trainable state	LoRA rank 16, alpha 16, dropout 0; original decision head retained; 165,041,156 trainable parameters in total.
Optimiser and rate	Fused AdamW; learning rate 2e-5 for head and adapter; linear decay, no warm-up, weight decay 0.
Batch and sequence	Training microbatch 1, accumulation 8; maximum 4,096 tokens, no truncation.
Objective and clipping	Hard-label cross-entropy; no label smoothing; global gradient-norm clipping at 1.0.
Run	Three epochs, 625 updates each; seed 20261104; one RTX 4090.
Final evaluation	Stock and tuned both native eager NF4/BF16, batch 4, identical option mapping.

A conventional adapter that teaches a language model to print a label would be a different experiment. This checkpoint includes the native head as well as the adapters. Fresh reload checks verified all 618 saved head-plus-adapter tensors. Unsloth supported the memory-efficient training workflow; native head preservation and classification loss were explicit implementation requirements.

Training batch, gradient accumulation and serving-runtime batch are different controls. The earlier LM Studio inference batch adjustment is not a training setting. Pinned revisions and stack identities appear on p. 17. R2; Clef model card.

Validation selects epoch three

All three checkpoints classified 499 of 500 validation examples correctly. Epoch three won the predeclared lower-negative-log-likelihood tie-break. There was no test-based search for the best epoch.

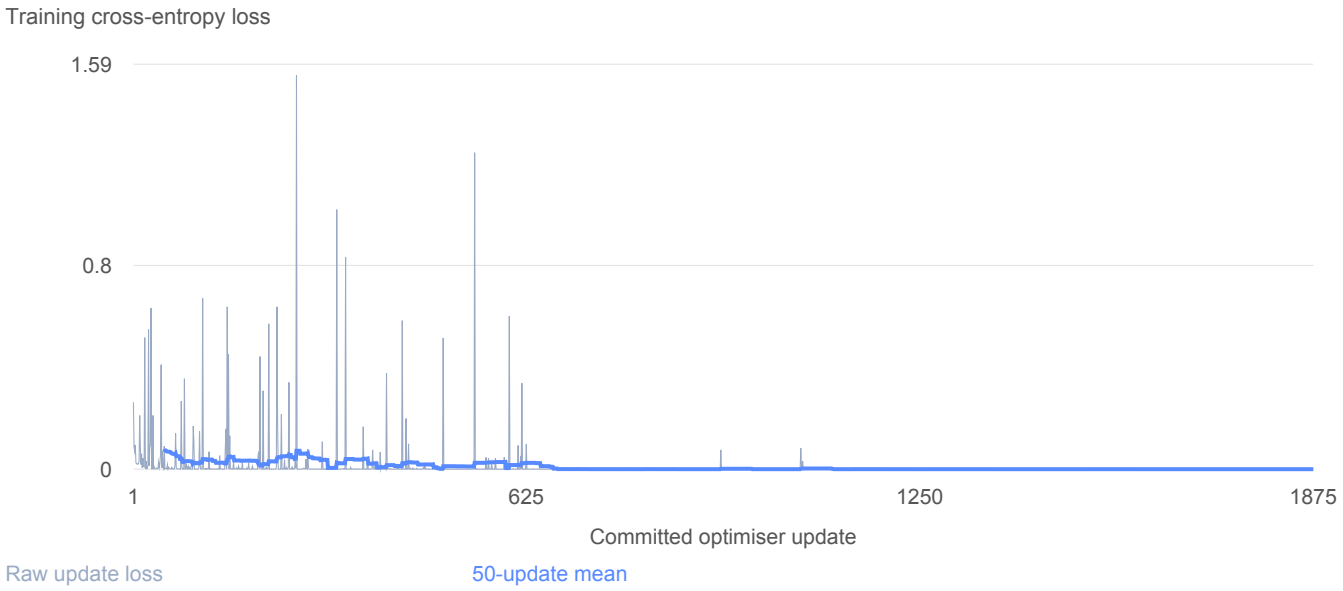


Figure 5. All 1,875 committed updates; latest retained attempt per step. The 75 replayed updates after recovery are omitted from this trajectory, not from the execution record.

Checkpoint	Update	Correct / 500	Accuracy	Validation NLL
Matched stock	0	488	97.60%	0.173600
Epoch 1	625	499	99.80%	0.006954
Epoch 2	1,250	499	99.80%	0.006027
Epoch 3, selected	1,875	499	99.80%	0.005593

The full production interval was 4.07 hours, including initialisation, validation, checkpoint handling and recovery. An interrupted save required resuming from update 625 and replaying 75 updates. The plotted trajectory uses the latest retained record for each of the 1,875 committed updates; the original journal preserves the replayed work.

Occasional loss spikes did not prevent validation improvement. Recorded gradient norms were measured before clipping, and the first 624 updates contained no non-finite losses or norms. The validation result is more informative than whether a small-batch training trace looks smooth.

NLL averages −ln p(reference class), so it can distinguish checkpoints that choose the same labels. Near-ceiling validation accuracy limits the resolution of this 500-row selector. R2.

The sealed test: 998 of 999 correct

Matched stock Clef classified 984 of the 999 sealed examples correctly; the selected checkpoint classified 998. Both returned valid native decisions throughout. Tuning corrected all fifteen stock mistakes and introduced one new mistake.

Same 999 examples	Stock correct	Stock wrong
Tuned correct	983	15 corrections
Tuned wrong	1 regression	0 both wrong

Figure 6. Paired sealed-test outcomes. Tuning corrected all 15 stock errors and introduced one new error; the net improvement is 14 examples.

Metric	Stock NF4	Fine-tuned NF4
Top-one accuracy	98.4985%	99.8999%
Errors	15	1
Macro F1, 50 classes	0.984775	0.998999
Negative log-likelihood	0.158402	0.003028
Invalid outputs	0	0
Batched eager forward total	270.446 s	303.003 s

The net gain is 1.40 percentage points, with a paired 95% bootstrap interval of +0.70 to +2.20 points. Exact two-sided McNemar p=0.000519. The error reduction is 93.3%; the absolute counts, fifteen errors becoming one, show the scale behind that percentage.

This is strong within-recipe generalisation. The sealed examples were separated from training and selection but share the taxonomy, authoring recipe and teacher process. The next comparison tests transfer to the earlier, differently constructed benchmark rather than projecting this near-perfect score onto arbitrary enterprise traffic.

Paired row bootstrap: 10,000 resamples, seed 20261104. Final exported test components are unique. Forward totals exclude the complete deployed client/server path and should not be read as request latency. R2.

The complete historical comparison

The table preserves every original comparator alongside the matched stock and fine-tuned Clef models. Every accuracy uses all 1,974 examples. Invalid results count as unsuccessful decisions; macro F1 averages fifty class F1 scores, with invalid outputs contributing a false negative to the true class.

Configuration	Correct	Invalid	Accuracy	Macro F1
Clef Q8 · Decisions API	1830	0	92.71%	0.9261
Jev 1.13 · Decisions API	1842	29	93.31%	0.9374
GPT-5.6 Luna · Responses · low	1774	103	89.87%	0.9207
GPT-5.6 Luna · Responses · high	1850	36	93.72%	0.9450
GPT-6 Luna · Responses · low	1840	60	93.21%	0.9447
GPT-6 Luna · Responses · high	1897	21	96.10%	0.9653
GPT-6 Luna · Decisions · numeric	1514	4	76.70%	0.7527
GPT-6 Luna · Decisions · descriptive / shuffled	1883	3	95.39%	0.9544
Clef stock NF4 · native eager	1827	0	92.55%	.9245
Clef tuned NF4 · native eager	1906	0	96.56%	.9651

The historical Q8 stock result, 1,830 correct, is close to the matched NF4 stock result, 1,827. The matched comparison is the cleaner test of the training effect because encoding, numerical format and scoring implementation are shared. Q8 remains visible to maintain continuity with the first paper.

Fine-tuned Clef's nine-example lead over Luna high is an observed result, not established superiority. Native output reliability contributes to the full-cohort ordering. On the 1,953 rows where both models produced valid outputs, Luna was correct on 1,897 and Clef on 1,885. Both the operational score and this conditional view belong in the interpretation.

Low/high settings apply to Responses, not native Decisions. "Descriptive / shuffled" changes label representation and uses a single frozen shuffled order. Other conditions have no applied shuffle. R1–R3.

Paired evidence supports the stock uplift

For each comparator, a paired comparison counts cases only tuned Clef gets right and cases only the comparator gets right. Complete scenario components are resampled, preserving the correlation among three variants of the same baseline scenario.

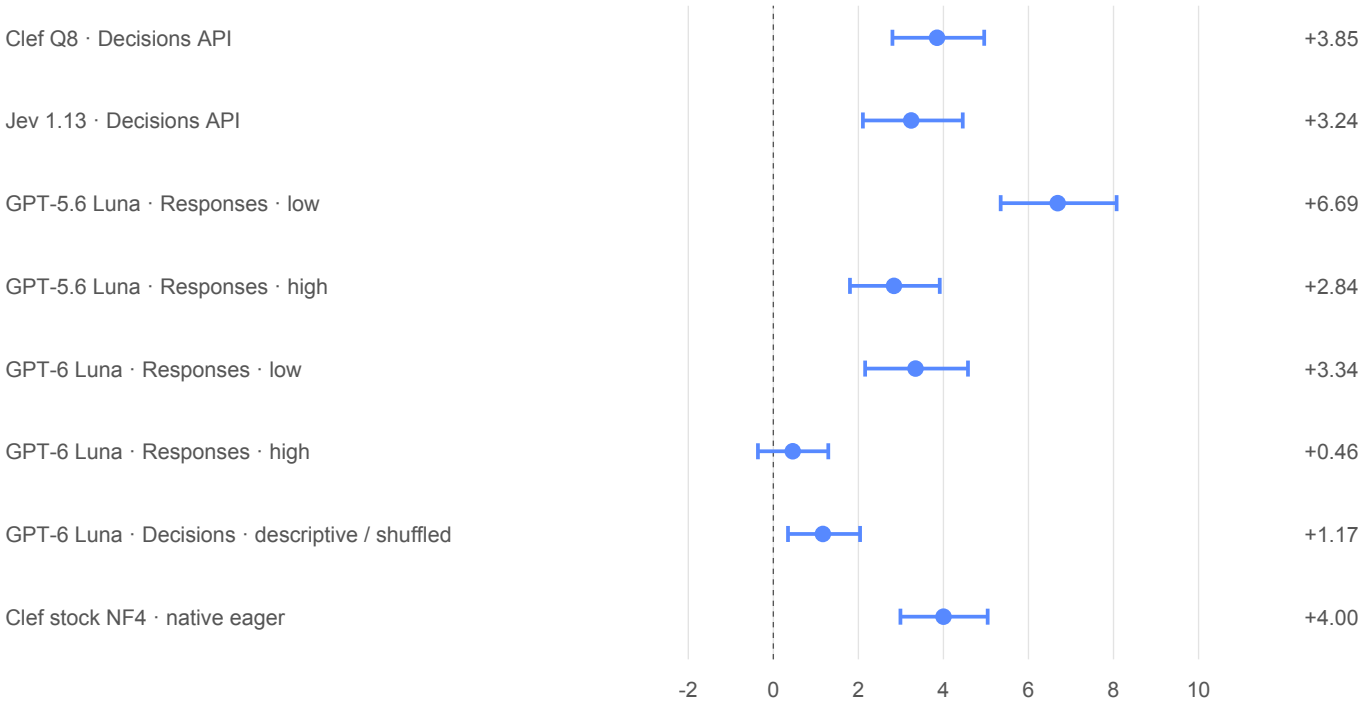


Figure 7. Fine-tuned Clef minus each comparator, in percentage points, using complete scenario families. An interval crossing zero does not establish a directional advantage. Numeric-label Decisions is tabulated separately because its gap is much larger.

Comparator	Tuned only	Other only	Net points	95% paired interval
Matched Clef stock NF4	86	7	+4.00	+2.99 to +5.04
GPT-6 Luna Responses high	39	30	+0.46	−0.36 to +1.29
GPT-6 Luna Decisions descriptive / shuffled	48	25	+1.17	+0.34 to +2.04
GPT-6 Luna Decisions numeric	411	19	+19.86	+17.50 to +22.30

The stock gain is supported and substantial. The strongest frontier comparison is much closer. An interval crossing zero leaves the direction unresolved; it does not demonstrate equivalence. The numeric-label Decisions configuration is retained as experimental history, not presented as the best way to use that interface.

Intervals are descriptive, conditional on recorded labels and cohort. No formal equivalence margin or multiplicity-adjusted superiority programme was predeclared. R3.

Workflow boundaries and usable outputs

The historical cohort combines 1,488 scenario variants with 486 deliberately constructed boundary cases. The selected Clef model performs particularly well on the scenario variants; Luna high retains a small observed advantage on the boundary subset.

Subset	Fine-tuned Clef NF4	GPT-6 Luna Responses high
Scenario variants, n=1,488	1,457 (97.92%)	1,442 (96.91%)
Boundary cases, n=486	449 (92.39%)	455 (93.62%)
Full cohort, n=1,974	1,906 (96.56%)	1,897 (96.10%)
Jointly valid rows, n=1,953	1,885 (96.52%)	1,897 (97.13%)

A request can concern the same document or transaction while asking for a different action. Extracting a record, approving a change and posting an authorised update are different workflows under this taxonomy. Recognising the small cue that changes the requested deliverable remains important even when aggregate accuracy is high.

Define success before comparing services

The benchmark requires a valid ranked decision under its declared output schema. A schema-invalid response is unsuccessful under that contract. An application that needs only a label, repairs outputs or retries failures would have different behaviour, latency and cost. Those alternatives should be specified before measuring their result.

The full-cohort score captures the usefulness of the returned output as specified. The jointly valid view isolates a narrower question about classification once both systems satisfy the output requirements. Clef's advantage in reliable native output should not be described as proof of superior semantic interpretation.

Class-level correct counts are retained in the appendix. With roughly forty examples per category, a one- or two-example movement should not be mistaken for a stable category-level ranking. Overall macro F1 improves from 0.9245 to 0.9651 in the matched Clef comparison.

The historical cohort is a development-used regression benchmark with provisional synthetic references. Its construction and boundary examples are documented in Paper 3. R1–R3.

Confidence improves, but needs calibration

Native Clef returns a full probability vector over the choices. Fine-tuning improves aggregate probability metrics, but concentrates confidence so strongly that some remaining errors become highly confident.

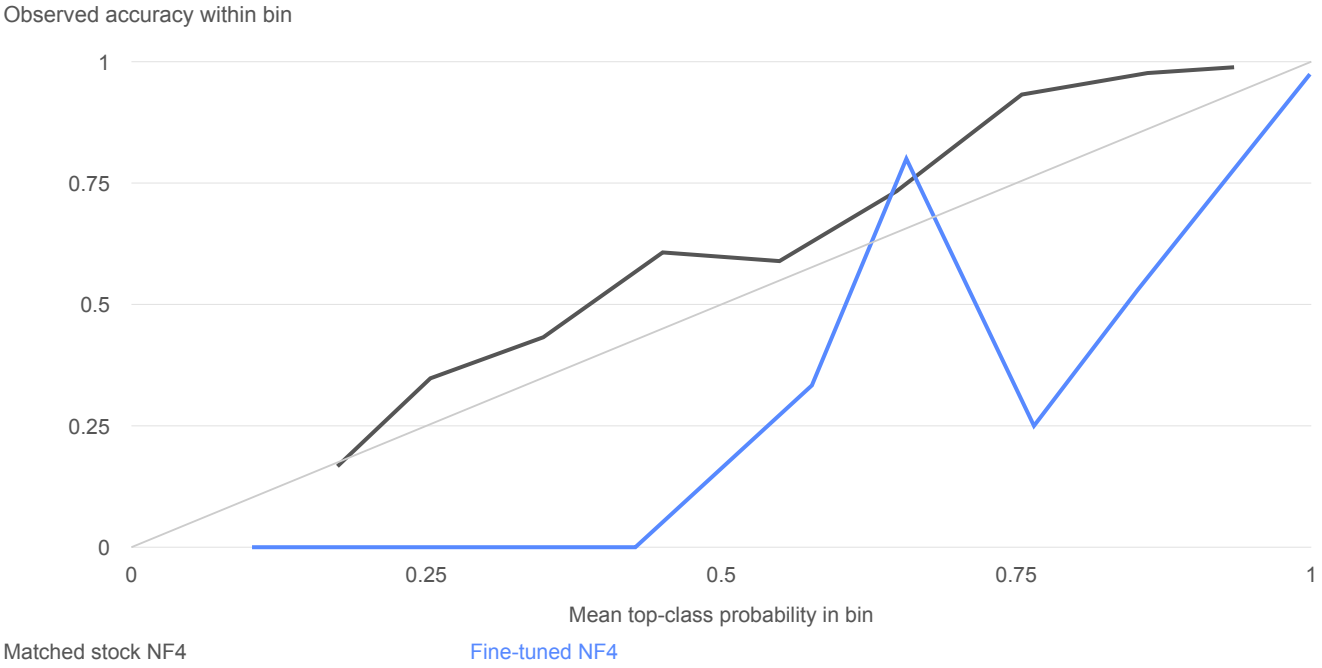


Figure 8. Historical reliability diagram using ten equal-width bins. Lines connect nonempty bins; low-support bins are noisy. The concentration of tuned predictions near one makes aggregate calibration scores insufficient for an escalation policy.

Historical metric	Stock NF4	Fine-tuned NF4
Top-three correct	1,925 / 1,974	1,951 / 1,974
Multiclass Brier score	0.122564	0.062183
Top-label ECE, ten equal bins	0.088047	0.029308
Incorrect decisions with confidence ≥0.95	5	43

Forty-three of the tuned model's 68 remaining errors have confidence of at least 0.95. Lower aggregate ECE is therefore not evidence that a high-confidence acceptance threshold is ready for production. The probability distribution becomes sharper as well as more accurate.

A separate representative calibration set is needed to design an escalation policy. This experiment did not fit a recalibrator or optimise abstention. Responses probabilities are partial self-reported ranked values, so full-vector NLL and Brier are not treated as directly comparable across those interfaces.

ECE is the bin-weighted gap between mean top probability and observed accuracy. Brier averages the sum of squared probability errors across fifty classes. R2–R3.

Local latency and matched forward timing

The first paper’s stock Q8 Clef endpoint had median request latency of 221 ms and p95 of 306 ms. The new fine-tuning comparisons use batched native eager forwards, which measure a different part of the system.

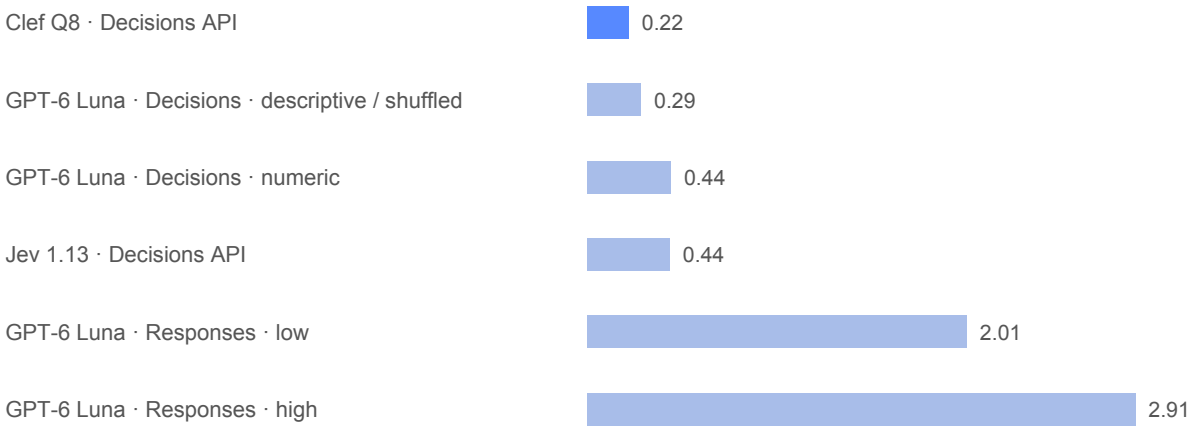


Figure 9. Historical median request latency in seconds, including terminal failures. Different serving and concurrency conditions apply. Native eager fine-tuning evaluation is deliberately excluded from this endpoint-latency chart.

Matched native evaluator, batch 4	Stock total	Tuned total
999 sealed examples	270.446 s	303.003 s
1,974 historical examples	517.639 s	574.278 s

The tuned historical forward total is approximately 10.9% higher than matched stock in this execution. It excludes model loading and the complete serving/client path. Dividing a batched total by the row count gives amortised forward time, not individual request latency.

A tuned deployment still needs a measured endpoint profile: warm and cold requests, concurrency, queueing, throughput and tail latency. Training microbatch, evaluation batch and serving-runtime batch are distinct settings. The earlier serving batch adjustment does not establish the correct training batch or tuned serving capacity.

Higher-end equipment may improve capacity, but no H200 Clef throughput was measured. The economic scenarios that follow therefore keep tuned serving performance and hypothetical rack capacity separate from observed endpoint measurements.

Historical chart includes terminal failures and differing client/concurrency conditions. It describes recorded systems, not a controlled hardware benchmark. R1–R2.

The cost of specialisation and hosted inference

Primary data production cost **$156.93**, reconciled to provider usage. The 4.07-hour training interval implies approximately **$2.80** of workstation depreciation and assumed electricity under the retained cost assumptions. Their **$159.73** subtotal excludes engineering, earlier pilots, final evaluation and ongoing serving operations.

Recorded configuration	Median s	p95 s	Run USD	USD / 1,000
Jev 1.13 · Decisions API	0.439	0.549	0.1747	0.0885
GPT-5.6 Luna · Responses · low	2.354	3.684	0.3415	0.1730
GPT-5.6 Luna · Responses · high	3.258	4.976	0.5310	0.2690
GPT-6 Luna · Responses · low	2.015	3.583	0.1764	0.0894
GPT-6 Luna · Responses · high	2.909	4.925	0.2962	0.1500
GPT-6 Luna · Decisions · numeric	0.445	0.691	0.3015	0.1527
GPT-6 Luna · Decisions · descriptive / shuffled	0.286	0.541	0.3363	0.1704

The API table covers the earlier 1,974-request workload, including failures. Costs are usage-priced at the recorded rates, with cache treatment preserved. They are not generic prices for every prompt length or reasoning configuration. Decisions values use exact usage-times-rate cost rather than rounded budget ceilings.

Primary data spend covers all 6,497 exported training, validation and test rows, plus failed, repaired and unused candidate work. It includes the generic pipeline’s probability-target pass, although Clef trained on hard labels. That is about 2.42 cents per exported example. The separate measured cost trial was $5.11; it is not included in the $159.73 narrow subtotal.

Local inference has no external per-call bill, but hardware ownership remains a real cost. The earlier Q8 run’s occupied-time depreciation and assumed electricity were approximately $0.095 for all 1,974 requests. The next pages include idle capacity and the broader ownership costs that this narrow run allocation omits.

Costs exclude tax and credit-purchase fees. Training energy is an assumed allocation, not a power-meter reading. Dataset production economics and route qualification are detailed in Paper 3. R1, R2, R4–R5.

Workstation ownership and utilisation

The retained workstation assumption is $15,000 capital cost, a three-year straight-line write-down and zero resale value. Maintenance, operator time, space/network and idle electricity are included in the fully allocated model.

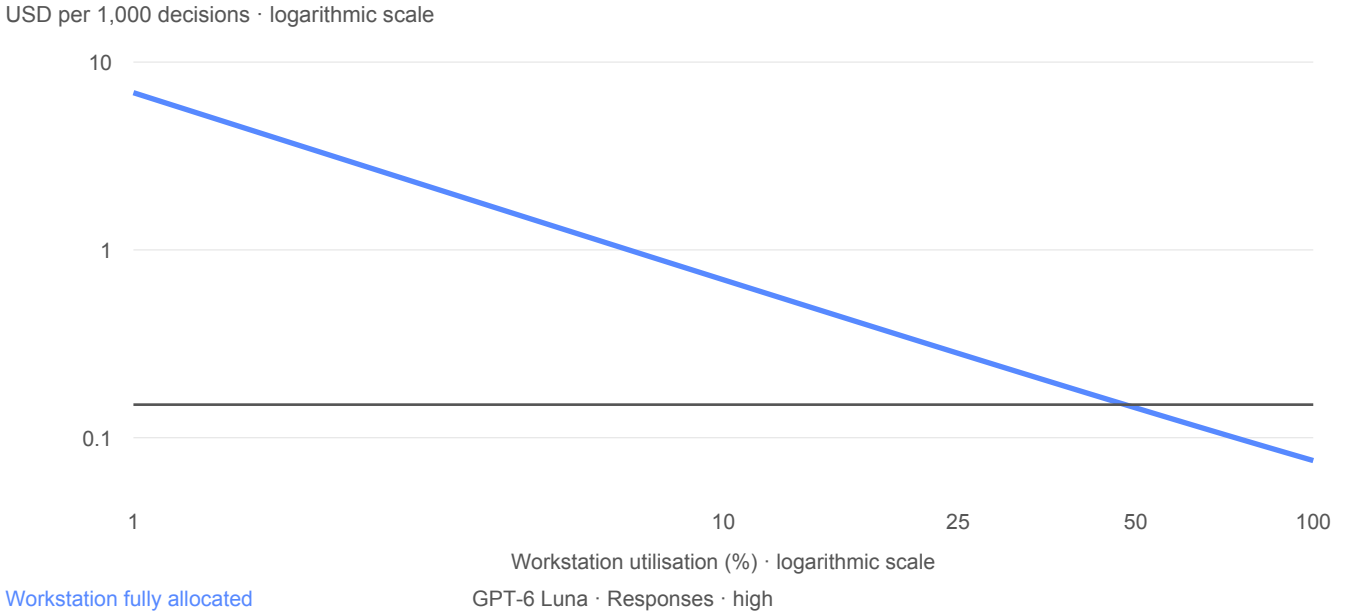


Figure 10. Retained paper-one stock Q8 capacity scenario, not a measured tuned-model serving curve. Hosted comparison uses the historical workload's usage-priced API cost and excludes client integration costs.

Assumption	Value
Annual depreciation	$5,000
Maintenance; operations; space/network	$750; $2,400; $300 per year
Idle / active power	0.10 / 0.65 kW
Electricity; facility multiplier	$0.15 / kWh; 1.20
Annual cost at utilisation u	$8,607.68 + $867.24 × u
Fully allocated USD / 1,000 at 1% / 10% / 50% / 100%	$6.873 / $0.694 / $0.144 / $0.076

The capacity assumption is the original stock Q8 serial measurement, 3.975 decisions/s while active. It is retained for continuity with Paper 1, not assigned to the tuned NF4 model. Its annual volume is 125,366,947 × u; actual tuned endpoint throughput must be measured before applying the same cost method.

At low utilisation, allocated ownership dominates the price of the decisions. Local deployment can still be valuable where context must remain within an approved environment. That sovereignty value is organisation-specific and is not given an invented dollar amount here.

Occupied-time training allocation: $2.324 depreciation plus $0.476 assumed facility-adjusted energy. Full service costs additionally include idle capacity, maintenance and operations. R5.

A transparent enterprise H200 scenario

The retained planning case is four eight-GPU H200 nodes: thirty-two GPUs in total, $400,000 per node plus $50,000 of infrastructure, written down over three years with zero resale value. The chart asks what average delivered throughput would justify that allocated cost.

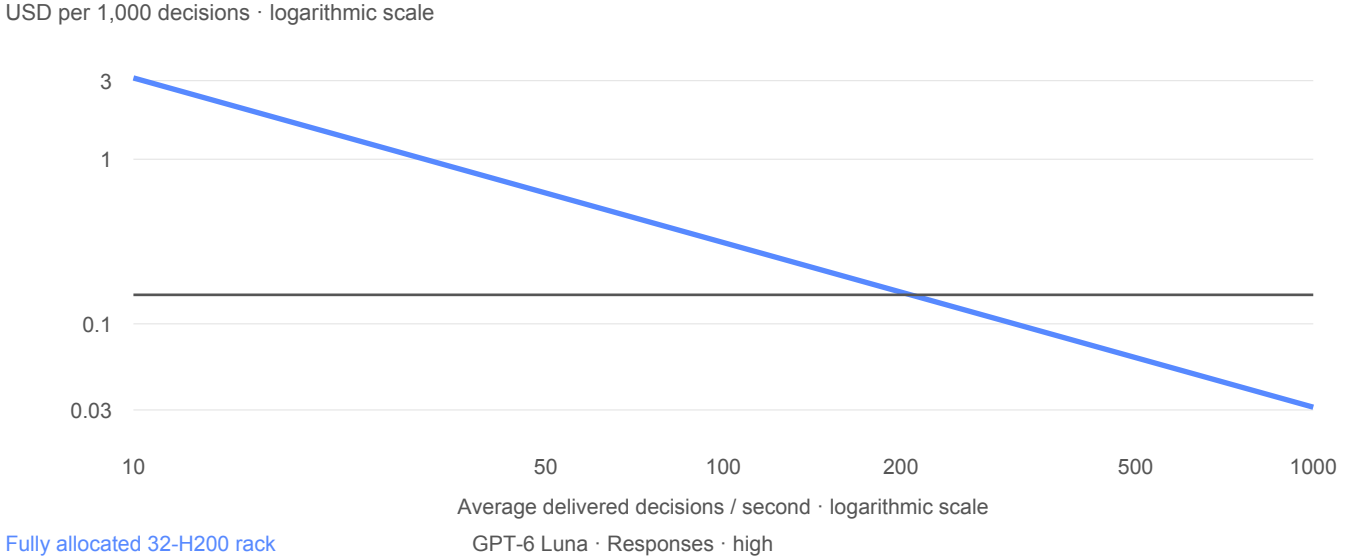


Figure 11. Required average delivered throughput, not a hardware benchmark. The full-rack cost crosses the historical Luna high API rate near 208 decisions/s, before any extra redundancy or disaster recovery.

Annual component	USD
Capital write-down	550,000
Node support plus ancillary maintenance	194,500
Colocation: 45 kW at $350 / kW / month	189,000
Operations: 0.25 FTE at $150,000	37,500
Connectivity, observability and backup	12,000
Total annual / three-year cost	983,000 / 2,949,000

At 100 average delivered decisions/s, the modelled cost is $0.312 per thousand; at 500/s it is $0.062. These are sustained average rates across the allocated rack, including idle time. They are not measured H200 capabilities. Power, cooling and space are included in colocation and are not billed twice in the model.

A shared rack should allocate both cost and capacity consistently. Extra high availability, spare capacity and disaster recovery are not assumed free. This workload need not occupy an entire rack; the example exposes the economics rather than recommending that purchase.

Planning assumptions inherited from Paper 1, not live procurement quotes. The cost crosses the historical Luna Responses high API rate near 208 average decisions/s. R5.

A capable enterprise specialist

The primary result is substantial: 4,998 training examples and a modest local run reduced Clef's historical benchmark errors by 53.7%. Its observed full-cohort accuracy reached 96.56%, compared with 96.10% for the strongest frontier configuration tested. A separate sealed test confirmed improvement within the generation distribution.

This is a practical route to frontier-range performance on a defined enterprise task. The small frontier difference establishes neither superiority nor formal equivalence; Luna remains more accurate among jointly valid outputs. Clef's reliable native interface and local handling of context are useful enterprise properties in their own right.

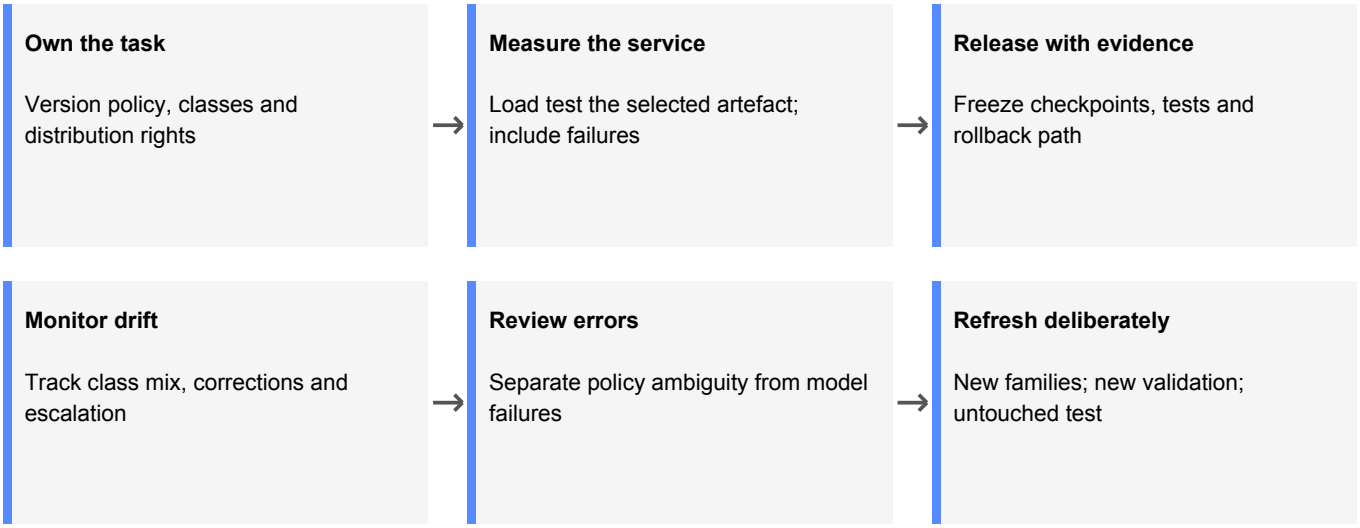


Figure 12. An enterprise operating cycle. Benchmark success is a release input; ongoing data, rights and service ownership determine whether it remains useful.

The narrow $159.73 data-and-training subtotal amortises to about $0.160 per thousand decisions over one million decisions, or $0.016 over ten million, before serving and maintenance. A hypothetical forty avoided errors per thousand could be worth much more than inference, but only if the benchmark uplift persists in actual traffic and the error consequences are understood. No client savings are claimed here.

The next deployment decision should use representative requests, measured tuned endpoint throughput, an explicit output contract and separately validated escalation thresholds. Model distribution and training-data rights also need to fit the intended use. Paper 3 explains the data and rights process behind this experiment.

This paper reports the primary training study. Original artefacts, prediction joins and checkpoint selection remain frozen. Dataset design and benchmark construction are separate contributions documented in Paper 3. R1–R5.

Methods and reproducibility

Accuracy counts valid correct top-one outputs over every scored example; invalid outputs are unsuccessful. **Macro F1** averages the fifty class-specific F1 scores, counting an invalid output as a false negative for its reference category. **NLL** averages −ln p(reference). **Brier** averages the sum of squared errors over all fifty class probabilities. **ECE** uses ten equal-width top-probability bins, weighted by the number of examples in each bin.

Historical accuracy intervals resample the 982 scenario components: 496 complete three-variant families and 486 singletons. Accuracy remains row-weighted within each resample. Marginal intervals use 5,000 replicates, NumPy PCG64 seed 20261102. The paired comparisons use the preserved component-bootstrap procedure and fixed input joins. The sealed test uses 10,000 paired row resamples, seed 20261104, with exact two-sided McNemar also reported.

Identity	Recorded value
Base model revision	fb4417d31c1d0f9e7e77f1141b40aaede7491e00
Unsloth revision	096fa323e9c0cfa7da58490f55daf6b06ce794e2
Training stack	Unsloth / Zoo 2026.11.1; PyTorch 2.11.0+cu130; Transformers 5.10.2; bitsandbytes 0.50.2; PEFT 0.21.2.
Selection	Epoch three, update 1,875; accuracy, lower NLL, earlier epoch on 500 validation rows only.
Saved trainable state	496 LoRA tensors + 122 original decision-head tensors; verified on fresh reload.
Numerical comparison	Matched native eager NF4/BF16 stock and tuned; batch 4; same encoder and option order.

The intervals are conditional on the recorded cohort and provisional references. They are not estimates from a probability sample of enterprise traffic. Family resampling addresses related examples, not prior development use or label uncertainty. No formal equivalence margin, repeated-seed analysis or multiplicity-adjusted superiority programme was specified.

The newer historical paired interval, +2.99 to +5.04 points, replaces an earlier row-bootstrap interval in the remote report. Recovery logs identify replayed work, and source hashes bind the frozen packs, selected checkpoint and evaluation inputs. No final test selected a checkpoint.

Detailed corpus construction, author/reviewer provenance, exclusion rules and file hashes are documented in Paper 3 and the preserved research evidence index. R1–R3.

Per-category results: categories 1–25

Correct counts on the same historical references. Stock and tuned Clef use matched NF4 native eager evaluation; Luna is GPT-6 Responses API with high reasoning. No shuffle is applied in these three conditions.

ID / sample category	n	Stock NF4	Tuned NF4	Luna high
1. Enterprise request intake and routing	40	40	40	40
2. Ticket triage and resolution	40	27	37	31
3. Policy-aware approval routing	40	38	38	39
4. SOP execution agent	40	35	35	35
5. Knowledge Q&A with citations	40	36	40	40
6. Research brief generation	40	39	39	40
7. Meeting-to-actions workflow	40	35	35	33
8. Email/thread action workflow	35	35	34	34
9. Document drafting workflow	40	37	40	38
10. Document redlining and comparison	40	40	40	40
11. Document extraction to structured data	40	40	40	40
12. Form completion and submission	34	30	33	34
13. System-of-record update workflow	40	33	34	36
14. Data reconciliation workflow	40	40	39	40
15. Master data maintenance	40	39	39	37
16. Exception investigation workflow	40	38	38	38
17. Escalation and handoff workflow	40	38	39	40
18. Status reporting workflow	40	40	40	40
19. Audit trail generation	40	38	39	40
20. Workflow monitoring and recovery	40	40	40	39
21. Customer support self-resolution	40	27	40	39
22. Technical support troubleshooting	40	33	38	38
23. Customer case prioritisation	40	38	39	38
24. Refund, return and credit handling	40	34	39	39
25. Complaint handling and de-escalation	40	40	40	38

Descriptive class counts, not separately powered superiority tests. The category names describe the ContextOne sample taxonomy and do not imply a production deployment. R1–R3.

Per-category results: categories 26–50

The near-balanced support exposes coverage, but one or two decisions can materially change an individual class percentage. Counts keep that resolution visible.

ID / sample category	n	Stock NF4	Tuned NF4	Luna high
26. Customer onboarding workflow	34	25	30	28
27. Customer health and churn intervention	40	38	40	39
28. Proactive incident communication	40	40	40	39
29. Multilingual customer workflow	40	37	38	39
30. Voice-of-customer mining	40	40	40	40
31. Lead enrichment and scoring	40	39	40	40
32. Account research and call prep	40	34	34	35
33. Sales outreach sequencing	40	40	40	40
34. RFP/RFI response workflow	40	40	40	40
35. Proposal generation workflow	35	35	35	35
36. CRM hygiene workflow	40	33	35	31
37. Quote and CPQ support	40	37	40	40
38. Discount and pricing approval	40	39	39	39
39. Renewal and upsell trigger workflow	40	40	40	39
40. Campaign planning and content workflow	40	36	37	40
41. Personalisation and segmentation workflow	40	39	40	40
42. Competitive intelligence monitoring	40	40	40	40
43. E-commerce recommendation workflow	40	28	39	39
44. Marketing compliance review	40	35	35	35
45. Code change implementation	40	37	39	40
46. Code review workflow	40	39	39	40
47. Test generation and maintenance	40	33	38	40
48. Build and CI/CD failure diagnosis	40	40	40	39
49. Production incident triage	40	38	38	40
50. Runbook execution workflow	36	35	35	34

Totals: stock NF4 1,827; tuned NF4 1,906; Luna Responses high 1,897, each out of 1,974. Earlier Q8 stock scored 1,830 and remains a separate condition. R1–R3.

Sources and the research series

R1 — Original benchmark and comparison. AI One, *Are open-source decision models useful for enterprises?*, 2 November 2026. Retained under output/clef-followon-paper-20261105/completed-study/paper-1.pdf, with paper-1-statistics and historical per-example comparator records. The TypeSafe run pilot-gpt6-luna-20260922 freezes the 1,974 public inputs, references, taxonomy, author/reviewer provenance and family mapping.

R2 — Primary training evidence. output/contextone-clef-production-20261104/training-results/clef-production-20261104-evidence-final.zip contains the frozen protocol, configuration, identities, journals, validation, selection, reload checks, sealed predictions and historical predictions. The recorded model is Cloudflare/Clef Flash. Unsloth's fine-tuning documentation provides general adapter/quantisation guidance; the native decision-head implementation and numerical outcomes are study evidence.

R3 — Paired historical analysis. completed-study/paired-comparisons.csv and comparison-method.json, under output/clef-followon-paper-20261105, preserve the nine tuned-versus-comparator comparisons. Immutable example IDs join all predictions to the same reference set. Complete native probability vectors support the derived calibration and per-category tables.

R4 — Data and recorded production spend. output/contextone-clef-production-20261104/final-billing.json, campaign.json, final-audit and frozen-pack/manifest.json. Actual export: 4,998 / 500 / 999. Provider-reconciled generation cost: $156.930187637. The measured earlier cost trial is separately recorded in docs/experiments/contextone-clef-v2/cost-trial-2026-11-04.md.

R5 — Ownership-cost assumptions. completed-study/paper-1-economics/economics.json and notes.md retain the workstation and H200 assumptions. Provider API costs use saved workload usage and rates. Training allocation adds the measured elapsed time from R2 to the retained depreciation and assumed energy model; it is not an invoice or a metered energy measurement.

Three papers, three questions

Paper 1 establishes stock-model capability, compares interfaces and examines decision-query design. **Paper 2**, this report, measures the gain from native task-specific training and its enterprise economics. **Paper 3**, *Building trustworthy enterprise benchmarks and synthetic training data*, explains the benchmark, recipe, review, provider rights and frozen-pack methods, including how Synthetic Data Studio supported the work.

Author: Mark J Sykes, AI One Labs. Prepared 6 November 2026. Master source register: docs/experiments/clef-followon-paper-evidence.md. Research records are retained; this report is not a blanket licence to redistribute underlying training data or artefacts.