Continuous fit and calibrate on arriving customer data

vega 4B and 0.8B, refitted and recalibrated as labelled rows arrive, against TypeSafe Jev 1.14 as shipped, tuned through its criteria, and behind the same calibration layer. Embedding Space technical report, 2026-11-06. vega measured on two T4s (4B) and one Apple M5 (0.8B); Jev on its live endpoint (model `jev-1.14.0`).

Summary

44.5%
typed-decisions: vega cold, no labels

73.7%
vega after 1,200 labelled rows, ECE 0.031

74.7%
Jev 1.14 zero-shot, ECE 0.047

60.1%
banking77: vega cold, no labels

87.5%
vega after 4,928 labelled rows, ECE 0.028

79.4%
Jev 1.14 zero-shot, ECE 0.078

What was measured. A deployment that only connects an API: labelled rows arrive over time, and every time a batch has landed the system refits one linear head per question on a frozen reader (0.8B or 4B here), recalibrates on rows the heads never saw, writes a new versioned calibration file and serves it. Nothing is trained; a refit is a closed-form fit over cached reader features and takes seconds. The hosted comparison is Jev 1.14, whose API has no fitting surface: its only lever is the text of the request, so the same labelled rows were written into its `criteria` as examples (the tuned arm), and its answers were also put behind the same temperature, conformal and review-band layer that vega uses (the layer arm), fitted on its own answers to the calibration rows.

Findings

- **typed-decisions.** Cold, vega 4B answers 44.5% of held-out decisions. After 1,200 labelled rows arrived through the loop it answers 73.7% at ECE 0.031, level with Jev as shipped (-1.0 points, inside the noise of 2,000 decisions), whose ECE is 0.047. The band serves 15.0% of decisions automatically at a measured precision of 0.943 against the 90% target. Writing the same rows into Jev's criteria lifts it to 75.1%, level with vega.
- **banking77.** Cold, vega 4B answers 60.1% of held-out decisions. After 4,928 labelled rows arrived through the loop it answers 87.5% at ECE 0.028, ahead of Jev as shipped by 8.0 points, whose ECE is 0.078. The band serves 92.5% of decisions automatically at a measured precision of 0.915 against the 90% target. Writing the same rows into Jev's criteria lifts it to 85.7%, still below vega.
- **The calibration split is not optional.** On typed-decisions, calibrating on the rows the heads were fitted on gives ECE 0.258 and routes 100.0% of decisions to auto; calibrating on held rows gives

ECE 0.031 at 15.0% auto with the precision holding. Same heads, same accuracy, opposite conclusions about what to automate.

- **What the two streams say together.** On a bespoke taxonomy with five typed questions and 1,200 rows, continuous fit on a frozen reader reaches a hosted decision model and beats it on calibration. On a 77-way intent stream with thousands of rows it keeps climbing past the hosted model while the band opens to most of the traffic. The lever that moved typed-decisions from 0.699 to parity was reader size, not more rows; the lever on banking77 was rows.
- **Cost.** Every Jev arm on both streams together cost under a dollar at the published price, so at benchmark scale cost is not the argument. The arguments are the calibration error, the band that says what to automate, a versioned calibration file per batch, and that the rows never leave the network.

The loop a customer connects to

The product side of this report is a loop, not a model. A customer sends labelled rows to one endpoint (a state, the questions, and the outcome once it is known) and reads decisions from another. Behind the endpoint:

1. **Buffer.** Rows accumulate per question. Nothing happens until each option of a question has enough rows to cross-validate and a calibration share is set aside (five per option and at least ten calibration rows, the same floor the open package enforces).
2. **Fit.** One linear head per question on the frozen reader's pooled features: PCA to 128 dimensions, then a multinomial logistic fit. Cross-validated against chance and against the engine's own accuracy on the same rows; a head that does not beat both is not served. Reader features are cached, so a refit reads only the new rows.
3. **Calibrate.** On the held rows the head never saw: a temperature by negative log-likelihood, a split-conformal threshold, and a review band at the precision target with a Wilson lower bound. Written as one JSON file stamped with the checksum of everything that changes an answer.
4. **Serve.** Every decision carries the calibration version it was made under, a confidence, a conformal set, an abstain flag and a route (auto or review).
5. **Watch.** The file records the answer mix and mean confidence it was fitted on; live traffic is compared against that reference and a drift past it triggers the next refit, which is the same loop again with more rows.

The small model in the loop is the reader: a frozen language model of 0.8B or 4B parameters that turns a state and a question into features. It never generates and it is never trained. Everything that adapts is closed form and takes seconds, which is what makes a refit per batch affordable and a human unnecessary. The rounds below are that loop run by hand on two public streams.

Setup

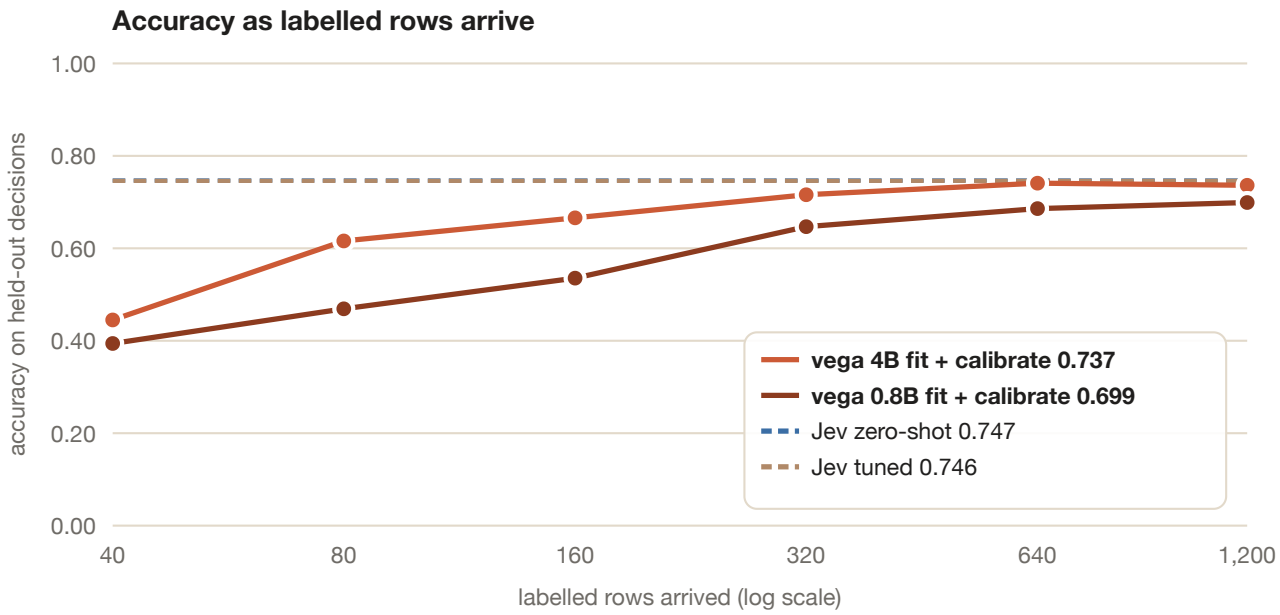
- **Streams.** typed-decisions (LocalLLaMA/typed-decisions): four workflows, five typed questions per case, 1,200 cases arrive interleaved across workflows, 400 held-out cases scored (2,000 decisions). banking77 (PolyAI/banking77): one 77-way intent question over customer messages, up to 4,928 rows arrive, 1,000 held-out rows scored, stratified by intent.

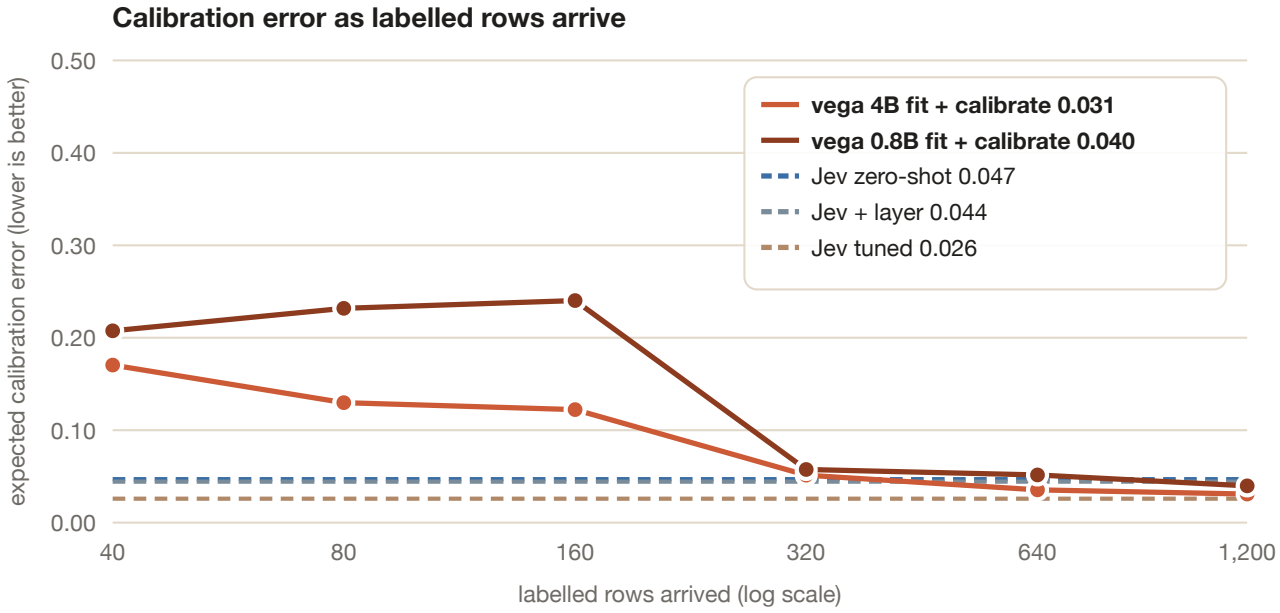
- **Rounds.** typed-decisions after 40, 80, 160, 320, 640, 1,200 arrivals; banking77 after 154, 308, 616, 1,232, 2,464, 4,928. At each round about a quarter of the arrived rows (by a hash of the id) are held for calibration and the rest fit the heads.
- **vega.** vegaml 0.15.0, Vega(mode='auto', heads=0, max_len=73728). `Vega.fit` keeps a head only when its cross-validated accuracy beats chance and serves it only when it beats the engine on the same rows. `calibrate` fits a temperature by NLL, a split-conformal threshold at alpha 0.1, and a review band at a 90% precision target with a Wilson lower bound; the file is stamped with its checksum.
- **Jev.** `jev-1.14.0`, pinned, one request per case, the dataset's own state and questions as the body (typed-decisions is published in exactly Jev's request shape). The tuned arm appends up to 2 labelled state excerpts per option to each option's criteria, from the rows that had arrived at that round. The layer arm fits our calibration on Jev's answers to at most 400 calibration rows.
- **Metrics.** Accuracy against the gold label; ECE over 15 bins of the reported confidence (max probability); Brier against the gold distribution where the dataset carries one; auto rate and precision inside the band; abstain rate from the conformal set. Jev reports no band, so its auto rate exists only behind the layer.

typed-decisions (4 workflows, 5 typed questions per case)

1,200 arrivals in the stream, 400 held-out cases scored at every round (2,000 decisions).

The curve





Cold, with no labels, vega answers 44.5% of decisions correctly at ECE 0.170. After 1,200 labelled rows it answers 73.7% at ECE 0.031. Jev 1.14 zero-shot answers 74.7% at ECE 0.047, and with the same labelled rows written into its criteria 74.6% at ECE 0.026.

Every round

arrived	fit rows	cal rows	heads serving	accuracy	ECE	Brier	auto	auto precision	calibration version	new reads
0 (cold)			0	0.445	0.170	0.297	n/a	n/a		
40	36	4	0 / 20	0.445	0.170	0.297	0.0%	n/a	4929bf82c353	200
80	69	11	13 / 20	0.616	0.130	0.247	0.0%	n/a	4929bf82c353	2439
160	128	32	18 / 20	0.666	0.122	0.233	0.0%	n/a	4929bf82c353	1330
320	242	78	20 / 20	0.716	0.051	0.149	7.8%	0.852	e8a7f1dbab0b	1664

arrived	fit rows	cal rows	heads serving	accuracy	ECE	Brier	auto	auto precision	calibration version	new reads
640	472	168	20 / 20	0.741	0.035	0.131	10.5%	0.972	0ff3e120e094	2751
1,200	899	301	20 / 20	0.737	0.031	0.144	15.0%	0.943	56d3f3dabb35	4940

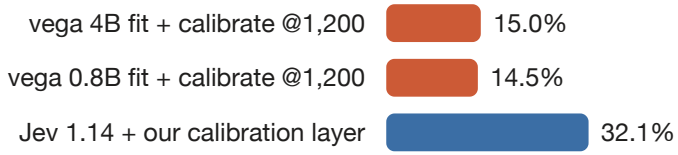
Heads serving is the number of questions whose fitted head beat the engine in cross-validation and so answers in production; the rest stay with the engine. The calibration version is the checksum stamped on the file written at that round. New reads are forward passes of the frozen reader that this round needed beyond what earlier rounds had read.

What a customer gets

arm	labelled rows used	accuracy	ECE	served automatically at 90% precision	precision inside the band	per-call cost	data leaves the network
Jev 1.14 as shipped	0	0.747	0.047	no band	n/a	$8 per million decisions	yes
vega 4B cold	0	0.445	0.170	no band	n/a	none	no
vega 4B, continuous fit + calibrate	320	0.716	0.051	7.8%	0.852	none	no
vega 4B, continuous fit + calibrate	640	0.741	0.035	10.5%	0.972	none	no
vega 4B, continuous fit + calibrate	1,200	0.737	0.031	15.0%	0.943	none	no
vega 0.8B cold	0	0.394	0.207	no band	n/a	none	no
vega 0.8B, continuous fit + calibrate	320	0.647	0.057	3.8%	0.763	none	no
vega 0.8B, continuous fit + calibrate	640	0.686	0.052	13.4%	0.918	none	no
vega 0.8B, continuous fit + calibrate	1,200	0.699	0.040	14.5%	0.935	none	no

Jev has no fit and no calibration endpoint, so as shipped it is the cold model at every row count; vega is the same frozen reader with a head and a calibration file that are refitted at every round. Per-call cost for Jev is its measured input tokens on this held-out set at the published $0.042 per million tokens, stated per million decisions.

Share of decisions served automatically at a 90% precision target



Precision inside the automatic band (target 0.90)



By question type

question type	n	vega acc	vega ECE	Jev acc	Jev ECE	Jev tuned acc	Jev tuned ECE
choice	600	0.732	0.074	0.715	0.044	0.700	0.076
noul	600	0.822	0.036	0.820	0.063	0.862	0.096
score	800	0.676	0.032	0.716	0.116	0.694	0.071

Ablation

arm (all from the same arrived rows)	accuracy	ECE	Brier	auto	auto precision	abstain
engine cold	0.445	0.170	0.297	n/a	n/a	55.0%
engine + calibrate, no heads	0.445	0.053	0.201	4.2%	0.807	0.0%
heads, no calibrate	0.737	0.124	0.215	n/a	n/a	n/a
heads + calibrate on the fitting rows	0.737	0.258	0.366	100.0%	0.737	32.7%
heads + calibrate on held rows (the product)	0.737	0.031	0.144	15.0%	0.943	1.1%
heads on every arrived row, no calibrate	0.753	0.117	0.213	n/a	n/a	n/a

What Jev can do with the same rows

arm	accuracy	ECE	auto band	cost for the held-out set	input tokens
Jev + examples in its criteria, 1,200 rows	0.746	0.026	no band	$0.0863	2,055,536
Jev + examples in its criteria, 80 rows	0.751	0.034	no band	$0.0782	1,861,236
Jev + our calibration layer (301 calibration rows)	0.746	0.044	32.1% at 0.946 precision	$0.0438	1,042,726

Its API has no fitting surface, so the labelled rows can only go into the request text: the tuned arm writes up to 2 state excerpts per option into each option's criteria. The layer arm leaves Jev's answers untouched and fits our temperature, conformal threshold and review band on its answers to the calibration rows, which is what gives it a band at all.

Calibration table, vega at the final round

confidence at or above	coverage	expected error	actual error
0.5	90.0%	0.201	0.232
0.6	76.6%	0.158	0.187
0.7	62.1%	0.113	0.151
0.8	49.1%	0.078	0.117
0.9	31.4%	0.038	0.078
0.95	20.6%	0.019	0.036

Calibration table, Jev zero-shot

confidence at or above	coverage	expected error	actual error
0.5	96.2%	0.204	0.240
0.6	83.5%	0.165	0.203
0.7	67.5%	0.121	0.174
0.8	52.1%	0.082	0.146
0.9	33.5%	0.043	0.100
0.95	20.5%	0.018	0.061

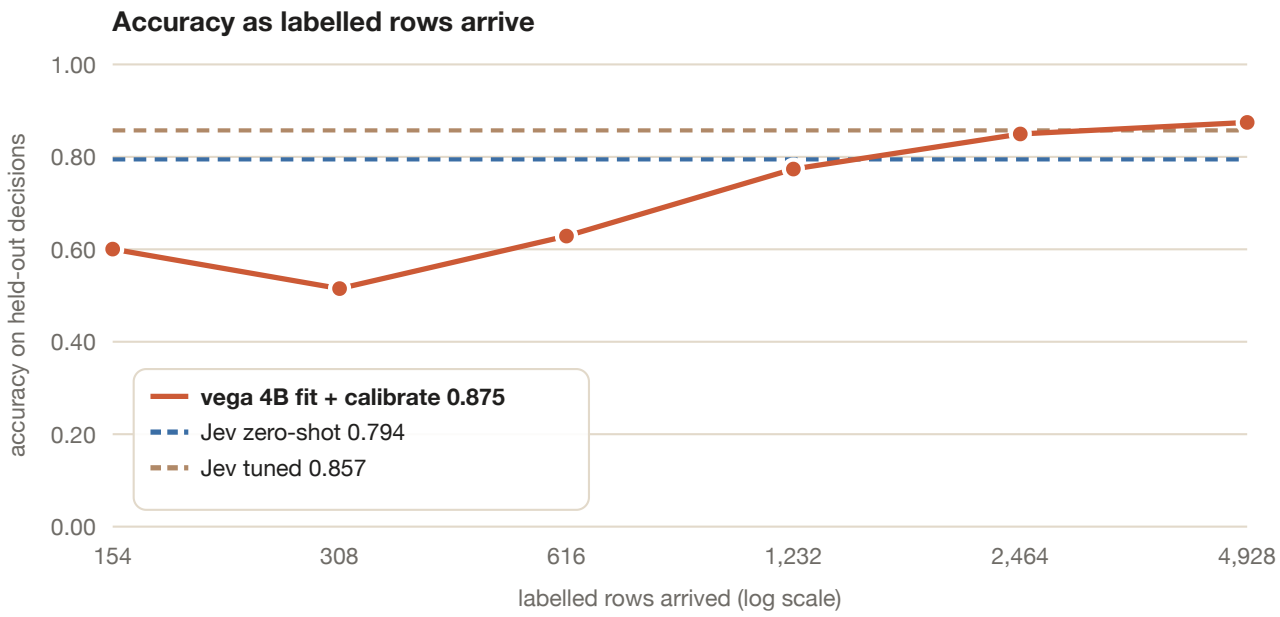
Calibration table, Jev tuned (tuned@1200)

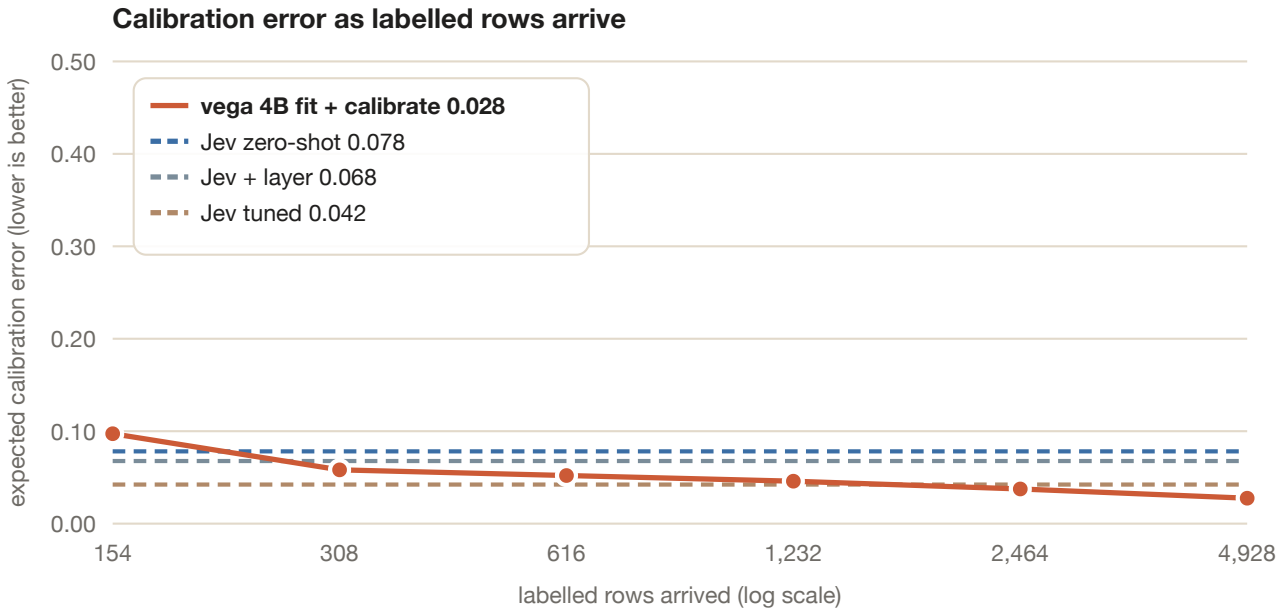
confidence at or above	coverage	expected error	actual error
0.5	94.8%	0.217	0.234
0.6	80.7%	0.176	0.204
0.7	65.5%	0.136	0.171
0.8	47.9%	0.091	0.122
0.9	27.6%	0.048	0.067
0.95	16.5%	0.025	0.030

banking77 (one 77-way intent question)

10,003 arrivals in the stream, 924 held-out cases scored at every round (924 decisions).

The curve





Cold, with no labels, vega answers 60.1% of decisions correctly at ECE 0.251. After 4,928 labelled rows it answers 87.5% at ECE 0.028. Jev 1.14 zero-shot answers 79.4% at ECE 0.078, and with the same labelled rows written into its criteria 85.7% at ECE 0.042.

Every round

arrived	fit rows	cal rows	heads serving	accuracy	ECE	Brier	auto	auto precision	calibration version	new reads
0 (cold)			0	0.601	0.251	0.636	n/a	n/a		
154	111	43	0 / 1	0.601	0.097	0.560	9.0%	0.964	a8287d37bc6d	154
308	226	82	1 / 1	0.515	0.058	0.642	1.3%	0.917	2d99aa44e586	1206
616	450	166	1 / 1	0.629	0.052	0.500	2.6%	1.000	1a39bd744151	610

arrived	fit rows	cal rows	heads serving	accuracy	ECE	Brier	auto	auto precision	calibration version	new reads
1,232	907	325	1 / 1	0.774	0.046	0.322	61.0%	0.922	ba22a2329504	1129
2,464	1833	631	1 / 1	0.850	0.037	0.230	86.7%	0.906	37b5a458f320	2165
4,928	3719	1209	1 / 1	0.875	0.028	0.185	92.5%	0.915	75d92fb547b6	4350

Heads serving is the number of questions whose fitted head beat the engine in cross-validation and so answers in production; the rest stay with the engine. The calibration version is the checksum stamped on the file written at that round. New reads are forward passes of the frozen reader that this round needed beyond what earlier rounds had read.

What a customer gets

arm	labelled rows used	accuracy	ECE	served automatically at 90% precision	precision inside the band	per-call cost	data leaves the network
Jev 1.14 as shipped	0	0.794	0.078	no band	n/a	$71 per million decisions	yes
vega 4B cold	0	0.601	0.251	no band	n/a	none	no
vega 4B, continuous fit + calibrate	1,232	0.774	0.046	61.0%	0.922	none	no
vega 4B, continuous fit + calibrate	2,464	0.850	0.037	86.7%	0.906	none	no
vega 4B, continuous fit + calibrate	4,928	0.875	0.028	92.5%	0.915	none	no

Jev has no fit and no calibration endpoint, so as shipped it is the cold model at every row count; vega is the same frozen reader with a head and a calibration file that are refitted at every round. Per-call cost for Jev is its measured input tokens on this held-out set at the published $0.042 per million tokens, stated per million decisions.

Share of decisions served automatically at a 90% precision target



Precision inside the automatic band (target 0.90)



By question type

question type	n	vega acc	vega ECE	Jev acc	Jev ECE	Jev tuned acc	Jev tuned ECE
choice	924	0.875	0.028	0.794	0.078	0.857	0.042

Ablation

arm (all from the same arrived rows)	accuracy	ECE	Brier	auto	auto precision	abstain
engine cold	0.601	0.251	0.636	n/a	n/a	79.7%
engine + calibrate, no heads	0.601	0.091	0.561	14.3%	0.939	0.0%
heads, no calibrate	0.875	0.243	0.265	n/a	n/a	n/a
heads + calibrate on the fitting rows	0.875	0.051	0.190	100.0%	0.875	17.9%
heads + calibrate on held rows (the product)	0.875	0.028	0.185	92.5%	0.915	1.7%
heads on every arrived row, no calibrate	0.885	0.245	0.255	n/a	n/a	n/a

What Jev can do with the same rows

arm	accuracy	ECE	auto band	cost for the held-out set	input tokens
Jev + examples in its criteria, 308 rows	0.857	0.042	no band	$0.1846	4,396,244
Jev + examples in its criteria, 4,928 rows	0.857	0.042	no band	$0.1899	4,521,908
Jev + our calibration layer (400 calibration rows)	0.794	0.068	64.9% at 0.940 precision	$0.1594	3,794,832

Its API has no fitting surface, so the labelled rows can only go into the request text: the tuned arm writes up to 2 state excerpts per option into each option's criteria. The layer arm leaves Jev's answers untouched and fits our temperature, conformal threshold and review band on its answers to the calibration rows, which is what gives it a band at all.

Calibration table, vega at the final round

confidence at or above	coverage	expected error	actual error
0.5	92.8%	0.083	0.085
0.6	88.6%	0.065	0.067
0.7	83.9%	0.049	0.059
0.8	78.8%	0.036	0.047
0.9	68.2%	0.020	0.030
0.95	58.7%	0.012	0.020

Calibration table, Jev zero-shot

confidence at or above	coverage	expected error	actual error
0.5	95.6%	0.106	0.178
0.6	88.5%	0.078	0.139
0.7	80.7%	0.051	0.111
0.8	73.8%	0.032	0.091
0.9	65.3%	0.016	0.066
0.95	58.8%	0.009	0.050

Calibration table, Jev tuned (tuned@4928)

confidence at or above	coverage	expected error	actual error
0.5	96.1%	0.080	0.118
0.6	91.2%	0.060	0.096
0.7	87.5%	0.048	0.079
0.8	81.1%	0.032	0.059
0.9	71.3%	0.015	0.040
0.95	64.7%	0.008	0.032

Cost and latency

stream	Jev arm	input tokens	cost at $0.042 per Mtok	requests
typed-decisions	zero-shot	378,236	$0.0159	400
typed-decisions	zero-shot + calibration layer	1,042,726	$0.0438	
typed-decisions	tuned@80	1,861,236	$0.0782	400
typed-decisions	tuned@1200	2,055,536	$0.0863	400
banking77	zero-shot	1,559,564	$0.0655	924
banking77	zero-shot + calibration layer	3,794,832	$0.1594	
banking77	tuned@308	4,396,244	$0.1846	924
banking77	tuned@4928	4,521,908	$0.1899	924

Jev live request latency from this machine: median 81.5 ms, p90 111.9 ms, p99 153.4 ms over 1,501 live calls. vega on the M5 reads one state for one question in about 65 ms; a refit reads only the rows that are new since the last round, so the cost of staying current is proportional to the arrival rate, not to the size of the history.

Reproduce

The run is a public Kaggle notebook: kaggle.com/code/nandakishorconvai/vega-fit. Accelerator GPU T4 x2, internet on, the Jev key in a secret named `JEV_API`; typed-decisions runs on GPU 0 and banking77 on GPU 1, and the output holds the same `results.json` and per-round calibration files this report is built from. The datasets are `LocalLLaMA/typed-decisions` and `PolyAI/banking77`.

```
JEV_API_KEY=... python harness.py --reader 4b     # the same run on any GPU, results in o
python build_report.py out/results.json          # this report
```

Every Jev response is cached in `out/jev_cache.jsonl` keyed by the request body, so a rerun scores the same answers at no cost. The calibration file written at each round is in `out/`.