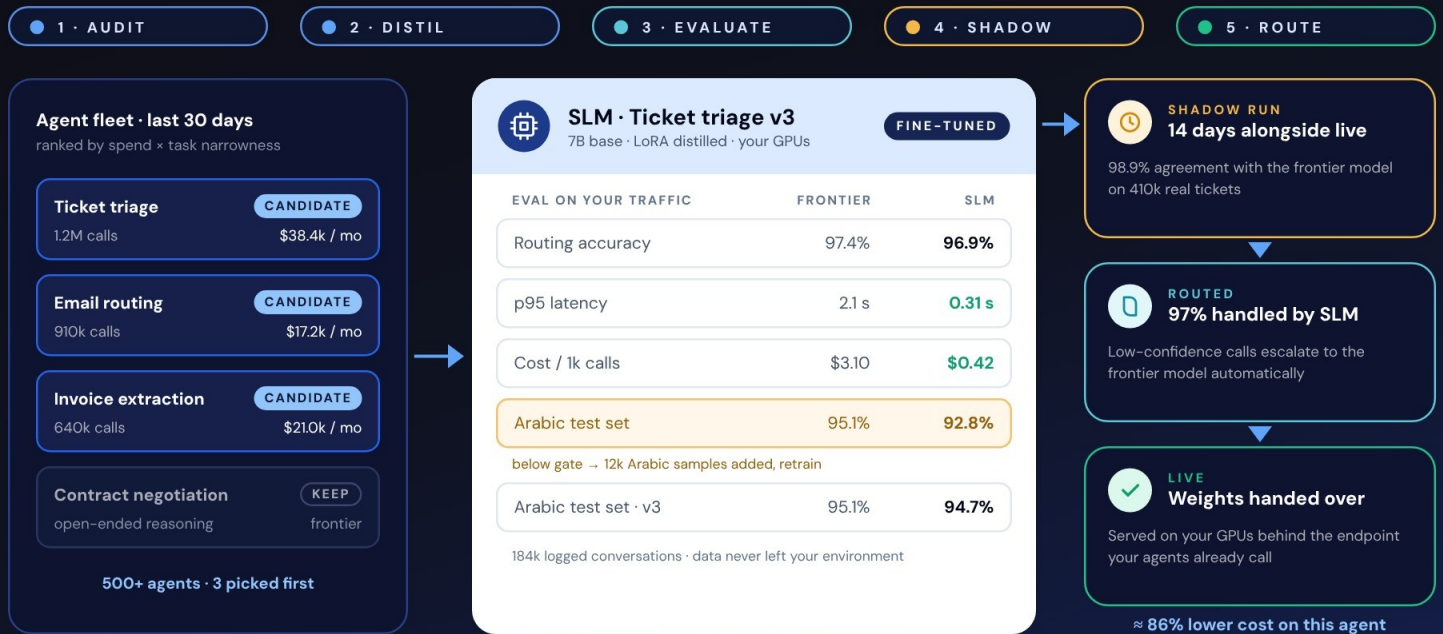


Your agents are built. Now make them cheaper, faster and yours.

At hundreds of agents in production, every frontier-model call adds cost, latency and a dependency you don't control. Most of those agents do narrow, repetitive work. **Evolvable.ai fine-tunes small language models for exactly that work:** trained on your own production logs, proven against your real traffic, running on your GPUs, with weights you own.



WHAT WE DELIVER

Start with three agents. Scale to the fleet.

Model engineering	Quality & proof	Serving & routing	Operations
Workload audit Rank every agent by spend, volume and task narrowness	Evals on your traffic Accuracy, tool calls and guardrails vs the frontier model	On-prem & air-gapped Served on your GPUs, private cloud or sealed network	Cost tracking Cost per task, before and after, in your ROI reporting
Distillation & fine-tuning Frontier behaviour, taught to a small model on your logs	Shadow runs The SLM runs beside live traffic before it answers anyone	Confidence routing Hard cases escalate to the frontier model automatically	Drift monitoring Retrain when your data or business changes
Arabic & bilingual tuning Gulf Arabic and English tested as separate gates	Release regression Every retrain re-runs the full suite before go-live	Sized for your hardware Quantised to fit the GPUs you already have	Full handover Weights, datasets and pipelines are yours

BUILT FOR	CTO & AI platform	CFO & finance	CISO & data office	Head of AI ops
	Lower spend without rebuilding a single agent	A cost per task you can forecast	Data and models stay inside your perimeter	Faster chains, fewer timeouts

60–80%

lower inference cost on migrated workloads

3–5×

faster responses on narrow, high-volume tasks

0

training records leave your environment

100%

of weights, datasets and pipelines owned by you

Four steps. *Models you own.*

01 Workload audit

Find the agents worth moving.

MONTHLY INFERENCE SPEND · TOP AGENTS

Ticket triage		\$38k
Invoice extraction		\$21k
Email routing		\$17k
Contract review		keep

3 candidates · narrow task, high volume, clean logs

- Pulls 30 days of call logs, cost and latency for every agent in your OS
- Scores each workload on volume, spend and how narrow the task is
- Delivers 3 ranked candidates with a business case per agent

2 weeks

to a ranked shortlist and business case

02 Distil & fine-tune

Your frontier model teaches your small one.



- Builds training sets from your own production conversations
- Masks PII before anything is trained, on your infrastructure
- Fine-tunes open models in English and Arabic, sized to your GPUs

6–8 wks

from logs to a benchmarked pilot model

03 Evaluate & shadow

Prove it before it answers anyone.

RELEASE GATES · v3 vs FRONTIER

✓ Accuracy -0.5 pt · gate 1 pt	✓ Tool-call trajectory 99.2% match
✓ Guardrail suite 0 new violations	✓ Shadow agreement 98.9% · 14 days

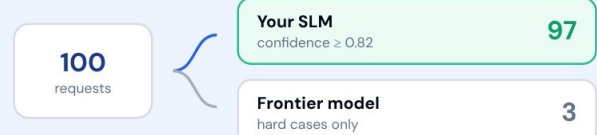
- Eval suite built on your real traffic: answers, tool calls, guardrails
- Hard release gates against the frontier model, per language
- Shadow runs beside live traffic before any cutover

≤ 1 pt

accuracy gap gate against the frontier model

04 Route & operate

Small model first. Frontier when it matters.



- Confidence-based routing sends hard cases to the frontier model
- Served behind the endpoint your agents already call: no rebuild
- Drift monitoring, scheduled retrains and regression on every release

≈ 85%

lower cost per task on routed workloads

ILLUSTRATIVE · 20 AGENTS
MIGRATED

What it adds up to

≈ \$1.1M

annual inference spend returned

≈ 4×

faster p95 response on migrated agents

20

agents on models you own

Figures are directional examples for a fleet of this size, not a guarantee. We baseline your real numbers in the audit.



You own the weights

Models, adapters, datasets and pipelines are handed over in full.



Data stays with you

Training and serving run on your infrastructure, air-gapped if needed.



Any open model

Qwen, Llama, Mistral, Phi, Gemma, Falcon, Jais. No lock-in.



Fits your OS

Drops in behind your existing model endpoint. No agent rebuild.

*Bring us your three costliest agents.
See the business case in two weeks.*

A two-week workload audit on your agent logs, on your infrastructure. You get three ranked candidates, target cost per task and a pilot plan.

BOOK A SESSION →

info@evolvable.ai · www.evolvable.ai