Trangile

Tech Driven Transformation

# Owning Your AI

## Private, Open-Source Intelligence

INDIA | UAE | HONGKONG | SINGAPORE | UK

# The shift in AI: Agentic Workflows

Trangile
Tech Driven Transformation

Enterprise token consumption gone up 13x in the last year, and most of that is agentic workflows that didn't exist an year ago

AI usage moved from chatbots to **agents** — autonomous systems that plan, act, and iterate

Agentic workflows consume **5–30× more tokens** than a single chat query

Every AI API call to the cloud is a metered charge


Your sensitive customer and company data leaves your walls on every call

The per-token price and consumption are rising much faster than physical resources can keep up with

Enterprises are blowing AI budgets months ahead of schedule

# SLM vs LLM

Trangile
Tech Driven Transformation

| | Large Language Models (LLMs) | Small Language Models (SLMs) |
|---|---|---|
| **Size** | 70B–1T+ parameters | Under ~10B parameters |
| **Strength** | Open-ended reasoning, broad knowledge, complex multi-step work | Focused, repetitive, narrow tasks |
| **Speed** | Slower, higher latency | Fast, often real-time |
| **Cost per task** | High | Low |
| **Best for** | "Help me figure this out" | "Do this specific thing, millions of times" |

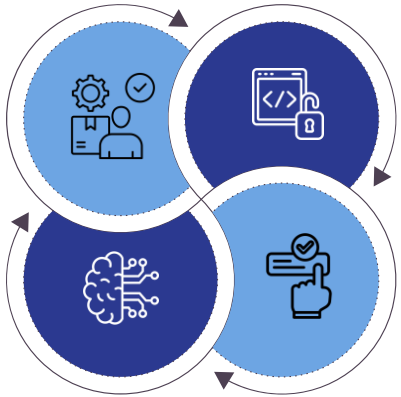
LLMs are **generalists**.
SLMs are **specialists**.

A fine-tuned SLM often **matches or beats** a frontier LLM on its specific task.

# Open source: don't rent it, own it

**Closed models** → you get an API endpoint. The model lives with the vendor.

**Own the intelligence, don't rent it**



**Open-source models** → you get the actual model file. You can host it, fine-tune it, own it.

This applies to *both* SLMs and LLMs

The best small and large models today are open source — Phi, Gemma, small Qwen variants, DeepSeek V4, Llama 4, the larger Qwens. You can pick the right size *and* keep ownership, depending on your need.

## The models are ready

The open-vs-closed quality gap has effectively closed for focused business tasks

Open-weight models now match or beat closed models on knowledge, math, and reasoning

This is no longer a quality compromise — it's just a deployment choice

# Why Open Source?

Trangile
Tech Driven Transformation

## Cost — break the meter, break the curve

- At high volume: **up to 5× cheaper** than premium APIs (e.g., $4.4K/mo vs $22.5K/mo at 500M tokens/day)

## Data Privacy — 100% data residency, by design

- **0% of your data** leaves your environment
- Eliminates exposure to cross-border transfer rules, No vendor breach, no third-party log

## Customization — specialists beat generalists

- Fine-tuned small models deliver **25–50% accuracy gains** on focused tasks vs. their longer versions
- Fine-tuned open models **outperform GPT-5 on 85%** of specialised enterprise tasks tested

## No vendor lock-in — your model is yours

- Frontier API prices have moved **±60% in a single year** — both up and down — and models get deprecated every 6–12 months
- Switch models in **one config change**, not a re-architecture

## Predictability — flat, known, budget-able

- Enterprise token consumption is up **~13× year-on-year** because of agentic workflows
- Owned-model cost: **a known number every month**, regardless of volume

# How we deliver: Our Approach

Trangile
Tech Driven Transformation

## What we do

Start from a proven open-weight model (no training from scratch — we *fine-tune*)

Fine-tune on your data with lightweight methods (LoRA / instruction tuning)

Deploy the model **inside your infrastructure** — your cloud tenant, your hardware

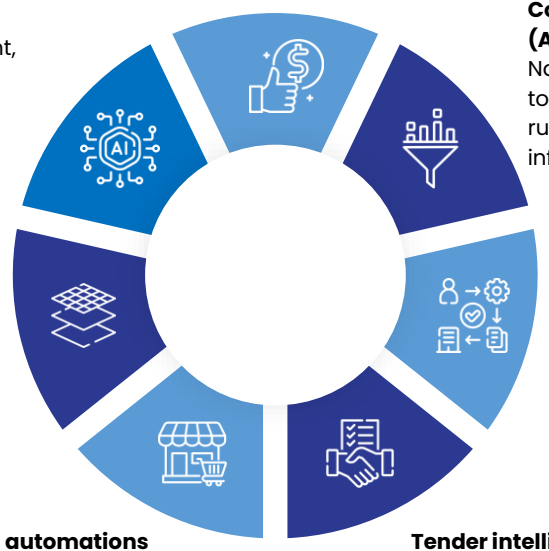
Optional: we operate and maintain it for you (managed service)

## What we've built recently

**Pricing intelligence agents**
Competitive market price tracking with confidence scoring

**AIRA**
AI Reconciliation Agent, in production — the exact pattern that fits COD reconciliation

**Conversational analytics (AIQ)**
Natural-language access to operational data, running on private infrastructure




**Agentic architecture**
For a GCC retailer — seven coordinated agents for allocation and replenishment

**Document processing & identity verification**
structured extraction from forms, IDs, and unstructured documents

**Retail batch automations**
Overnight POS uploads and inventory refresh

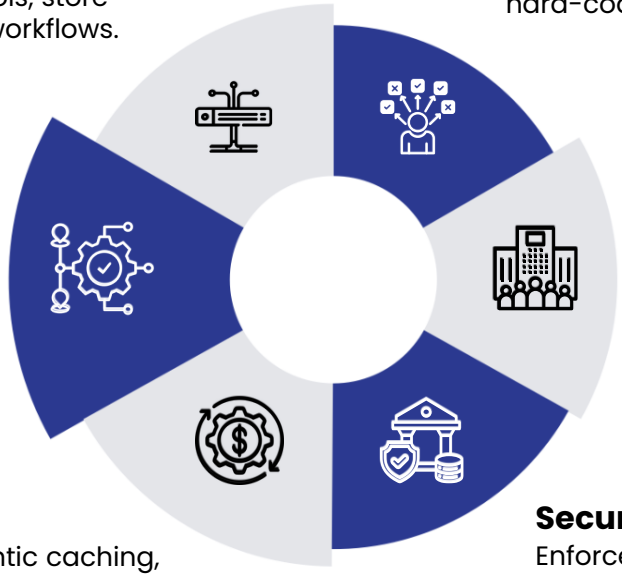
**Tender intelligence systems**
Auto-analysis of 100+ page RFPs, gap detection, bid generation

# AI Model Farm: Tying It Together

Trangile
Tech Driven Transformation

**One AI Gateway**
A single internal API and portal for employees, apps, agents, BI tools, store systems, and contact-center workflows.

**Model Choice without Chaos**
Lets teams test and fine-tuned models without hard-coding vendors into applications.



**Intelligent Routing**
Automatically chooses the best model based on task type, data sensitivity, user entitlement, latency, model quality, and budget.

**Enterprise Knowledge**
Grounds answers in approved data: product catalog, SOPs, HR policies, inventory, pricing, merchandising, tickets, and training content.

**Cost Optimization**
Reduces token spend through semantic caching, prompt compression, smaller models, reusable templates, and provider price comparisons.

**Security & Governance**
Enforces SSO, RBAC, DLP, PII masking, data residency, audit trails, model approvals, and policy-based external calls.

Trangile
Tech Driven Transformation

# Working Hard To Grow Your Business Everyday

For More Information:
Contact Us
www.trangile.com
Email: marketing@trangile.com