Sovereign AI Stack — one-page brief

A private LLM that physically can't leak: self-hosted multimodal serving where "client data never touches a cloud API" is enforced by routing, not policy. Public repo: github.com/MushiSenpai/mushishi-sovereign-ai-stack

What it is

- **Routing-enforced privacy tiers** — the "client" profile refuses to fall back to cloud at all; it fails loudly instead of degrading silently. Auditable, because the whole configuration is public.
- vLLM + Nemotron (NVFP4) on one RTX 5090; always-on CPU fallback (llama.cpp) so agents survive GPU mode-switches; optional cloud budget layer for the tiers your data rules allow.
- Operations included: default-deny firewall, Tailscale-only exposure, monitored 3-2-1 backups, runbook handover.

Measured (not projected)

METRIC	MEASURED	NOTE
Single-stream throughput	`276 tok/s`	real multimodal load
Context window	`180K tokens`	FP8 KV cache, ~28–30 GB of 32 GB
Concurrency knee	`~8 concurrent`	~2.65× aggregate; past it, users just wait
Practical users / card	`10–15 heavy`	~30–40 light chat users, <2.5 s latency

> **What it isn't:** a 30B local model is not a frontier model. The stack is honest about that — sensitive work stays local by architecture; anything routed to cloud is a tier your data rules explicitly allow. The decision log (including six hours of failed TensorRT-LLM debugging) is published.

Engagement

- **AI Readiness Audit** — SGD 500–1,000: two hours + a written gap analysis and roadmap. The low-commitment first step.
- **Deployment** — SGD 3,000–8,000 by scope: ~2 weeks spec + staging, delivered remotely onto your hardware; on-site sprints available for overseas engagements. Hardware spec'd to buy or deployed on yours.
- **Health retainer** — SGD 800–1,500/mo: backups verified, watchdog reviewed, dependencies current.

Madhan Kumar Reddy · Singapore · reddy.madhankumar.sg@gmail.com · theinvalid.me — every number above is a measured run on one RTX 5090 (32 GB); benchmarks and failure logs are public in the repo. Generated 2026-07-02.