

Sovereign AI Stack 2026

GPU-Native Execution Strategy for Enterprise AI Sovereignty

Subtitle: Moving beyond external AI API dependence to run the latest AI models, RAG, agents and LLMOps on dedicated enterprise GPU infrastructure

Published by: gcube.ai

Document version: v0.2 rebranding revision

Date: 2026-07-02

Intended audience: CIOs, CTOs, CDOs, AI transformation leaders, security/infrastructure leads, business automation leaders

Core positioning: gcube.ai is a GPU-Native AI Runtime that lets enterprises run the latest AI models, agents, RAG and LLMOps within their own data, security and GPU infrastructure.

Cover Copy

Move from an enterprise that calls external APIs to an enterprise that runs its own AI services.

In 2026, enterprise AI competitiveness depends not on which model you use, but on how safely and efficiently you can run AI within your own data, security and infrastructure. This whitepaper presents the Sovereign AI Stack framework and a GPU-Native execution architecture for applying the latest AI capabilities to enterprise work.

1. Executive Brief

The first wave of generative AI began with external APIs and SaaS tools. Many enterprises quickly adopted ChatGPT, Copilot, image generation tools and AI search services and saw their productivity potential. But the moment a PoC scales into real operations, enterprises hit new bottlenecks.

- External API token costs grow with usage.
- Customer data, internal documents, source code, and medical and financial information are sent outside the company.
- General-purpose models don't sufficiently understand internal terminology, policies and business context.
- AI answer quality, hallucination, latency and cost aren't measured continuously.
- As departmental PoCs multiply, models, data, permissions and GPU resources become fragmented.

The key shift for enterprise AI in 2026 is therefore moving from consuming AI APIs to operating a private AI runtime. Enterprises must not simply call the latest AI models; they must run them within their own data, permission, security and evaluation systems.

The Sovereign AI Stack described in this whitepaper is not a “collection of open-source tools.” It is a complete operating system for enterprises to regain control over how their AI runs.

Sovereign AI Stack =
 Open-Weight Frontier Models
+ Private AI Gateway
+ Enterprise Knowledge AI
+ Agentic Workflow Engine
+ AI QualityOps
+ Trust & Governance Layer
+ GPU-Native Runtime

The role of gcube.ai

gcube.ai is a GPU-Native AI Runtime that helps enterprises carry the latest AI capabilities past the experimental stage and into real productivity, cost savings, stronger security and new services.

2. The Enterprise AI Shift

2.1 Five bottlenecks of enterprise AI in 2026

Bottleneck	Current symptoms	Solution direction from the Sovereign AI Stack perspective
Data security	Hard to send sensitive data to external APIs	Private AI Gateway, on-premise/private deployment, PII masking
Cost predictability	Rising token charges, overlapping SaaS licenses	GPU-Native Runtime, model routing, caching, quota
Domain accuracy	Lacks internal terminology, policy and document context	Domain-Adaptive Intelligence, Enterprise Knowledge AI
Operational quality	Can't measure quality, hallucination or latency after deployment	AI QualityOps, tracing, eval, regression test
Scalability	Departmental PoCs don't connect into a platform	Enterprise AI Operating Platform, ModelOps, governance

2.2 Not “open source” but “Enterprise-Deployable AI”

The implementation stack in this whitepaper is built on an open ecosystem including vLLM, SGLang, LlamaIndex, LangGraph, ComfyUI, Phoenix, Presidio and KServe. However, the headline message is defined not as “open source” but in the following language.

Common term	Whitepaper term
Open-source AI	Open AI Ecosystem
Open-source models	Open-Weight Frontier Models
Self-built	Private AI Runtime
In-house ChatGPT	Private AI Gateway
RAG chatbot	Enterprise Knowledge AI
AI agent	Agentic Workflow Engine
Model evaluation	AI QualityOps
Security filter	Trust & Governance Layer
GPU server	GPU-Native AI Runtime

3. Sovereign AI Stack Framework

[Diagram: Diagram 1 - Sovereign AI Stack — diagrams/01-sovereign-ai-stack.png]

3.1 7-Layer Framework

Layer	Role	Key technologies
Application Layer	The touchpoint where employees, customers and business apps use AI	AI Portal, Chat UI, API Client, Workflow UI
Agent Layer	The execution layer that plans toward goals and calls tools	LangGraph, CrewAI, AutoGen, MCP
Knowledge Layer	Turns internal documents and data into searchable knowledge	LlamaIndex, Haystack, Milvus, Qdrant, Weaviate
Model Layer	The model layer that understands, generates and analyzes enterprise work	Llama, Qwen, Mistral, Gemma, DeepSeek families
QualityOps Layer	Operates quality, cost, latency and regression testing	Phoenix, Langfuse, Ragas, Promptfoo
Trust & Governance Layer	Privacy, permissions, auditing and safety policies	Presidio, NeMo Guardrails, OPA, LLM Guard
GPU-Native Runtime	Runs training, inference, documents, video and agents	gcube.ai GPU Cloud, vLLM, SGLang, KServe

3.2 Core principles

1. Private by Design: Process sensitive data inside the enterprise's controlled environment wherever possible.
2. Composable by Architecture: Design models, RAG, agents, evaluation and security as replaceable layers.
3. GPU-Native by Operation: Operate based on the GPU characteristics and cost structure of each AI workload.
4. QualityOps by Default: Continuously measure quality, cost and safety after deployment.
5. Human-Governed Autonomy: Agents run automatically, but high-risk actions require human approval and leave audit logs.

4. 10 Strategic AI Workloads for 2026

4.1 Domain-Adaptive Intelligence

Dedicated language intelligence that understands enterprise data

Conventional term: LLM fine-tuning

Strategic meaning: The layer that adapts general-purpose models to the enterprise's language, policies and domain judgment criteria

Enterprise problem

General-purpose LLMs are strong on general knowledge but don't accurately reflect internal terminology, product names, support policies, or legal, medical and manufacturing context. Prompts alone make it hard to reliably control answer style, compliance and use of specialist terms.

Latest AI capabilities

- LoRA/QLoRA-based parameter-efficient fine-tuning
- DPO/ORPO-based preference alignment
- Separate operation of instruction datasets and preference datasets
- Tuning on Korean and multilingual domain datasets
- Deployment after regression tests against evaluation sets

Implementation stack

- Unsloth: fast LoRA/QLoRA training
- LLaMA Factory: model tuning via WebUI/CLI
- Axolotl: large-scale experiments and fine-grained training configuration
- PEFT, TRL, bitsandbytes, FlashAttention
- Serving tuned models with vLLM/SGLang

Deep Case Scenario — Domain-specific customer support LLM

Item	Details
Industry	Commerce, finance and telecom customer centers
Enterprise problem	Agents handle repetitive inquiries every day, but general-purpose chatbots don't reliably follow company policies, product names or prohibited expressions
Current approach	FAQ chatbot + manual handling by agents. Rules must be updated whenever policies change
How AI is applied	Convert top-quality support logs, FAQs, refund/exchange policies and prohibited-expression guides into instruction/preference datasets, then tune with LoRA/DPO
Tech stack	Presidio de-identification → Argilla review → Unsloth/LLaMA Factory tuning → Promptfoo evaluation → vLLM serving
GPU workload	Data cleaning on mixed CPU/GPU, LoRA training on A100/H100-class, production inference on L40S/A100/H100-class
Expected KPIs	30–60% shorter average response drafting time, fewer repeat inquiries, more consistent support quality, lower external API costs
gcube.ai point	Integrated training, evaluation and serving in a dedicated GPU environment, without sending sensitive support data outside

gcube.ai message

gcube.ai helps enterprises build dedicated language intelligence from their own data and run it safely on a private AI runtime instead of external APIs.

4.2 Private AI Gateway

An in-house AI API layer that replaces external API dependence

[Diagram: Diagram 2 - Private AI Gateway — diagrams/02-private-ai-gateway.png]

Enterprise problem

External LLM APIs are quick to start with, but cost, security and SLA problems grow with usage. Enterprises need to unify diverse models behind a single API and manage permissions, cost, logs and quality centrally.

Latest AI capabilities

- OpenAI-compatible API
- continuous batching, prefix caching, KV cache optimization
- speculative decoding
- FP8/NVFP4 and other quantization optimized for the latest GPUs
- Model routing and fallback
- API quota, chargeback, tenant isolation

Implementation stack

- vLLM, SGLang, TensorRT-LLM, TGI
- LiteLLM or an API gateway
- Prometheus/Grafana, OpenTelemetry
- Phoenix/Langfuse tracing

Deep Case Scenario — Private AI Gateway for a financial institution

Item	Details
Industry	Finance and insurance
Enterprise problem	Employees use external LLMs individually, creating risks of entering customer information and problems managing cost
Current approach	SaaS AI accounts per department; usage and data movement are hard to track
How AI is applied	Build an in-house standard AI gateway that routes every LLM request to dedicated models through permission, masking, logging and cost policies
Tech stack	API Gateway/LiteLLM → Presidio → vLLM/SGLang → Phoenix/Langfuse → Grafana
GPU workload	7B–14B models on L40S/A100, 32B–70B models with A100/H100 tensor parallelism, high-performance inference on H100/H200/B200-class
Expected KPIs	Fewer external API calls, cost visibility by department, sensitive data kept from leaving, p95 latency under control
gcube.ai point	Operated as the enterprise's internal standard AI infrastructure by combining a dedicated GPU pool with an API layer

gcube.ai message

Move away from uncontrolled calls to external AI APIs and build an enterprise-dedicated Private AI Gateway on gcube.ai.

4.3 Agentic Workflow Engine

An autonomous work engine that judges and acts

[Diagram: Diagram 4 - Agentic Workflow Engine — diagrams/04-agentic-workflow-engine.png]

Enterprise problem

RPA is strong at fixed procedures but weak at exceptions, natural-language judgment, document understanding and tool selection. Enterprises need to automate search, drafting, review, system entry and approval requests as a single unit of work.

Latest AI capabilities

- stateful agent graph
- function calling / tool use
- MCP-based tool connections
- multi-agent orchestration
- human-in-the-loop approval
- traceable action log

Implementation stack

- LangGraph, CrewAI, AutoGen
- MCP, Playwright, internal API tools
- SGLang/vLLM tool-use model serving
- Phoenix/Langfuse agent trace

Deep Case Scenario — B2B sales proposal agent

Item	Details
Industry	B2B SaaS, manufacturing, IT services
Enterprise problem	Sales reps spend a lot of time on CRM lookups, customer research, proposal drafts and follow-up emails
Current approach	People draft manually, switching between CRM, the web and document templates
How AI is applied	The agent queries the CRM, researches

	industry pain points, pulls in internal proposal templates and success stories via RAG, and drafts the proposal. A human approves before sending
Tech stack	LangGraph planner → CRM API tool → Enterprise Knowledge AI → LLM draft → human approval → audit log
GPU workload	Many short LLM and RAG calls. Autoscaling on L40S/A100/H100 as concurrent users grow
Expected KPIs	50–80% shorter proposal drafting time, shorter follow-up email lead time, higher sales productivity
gcube.ai point	Reliably handles the inference, search and document generation workloads that arise while agents run, on a dedicated GPU runtime

gcube.ai message

gcube.ai provides the GPU-Native execution environment AI agents need to connect to business systems and judge, draft and act.

4.4 Enterprise Knowledge AI

Turning internal documents into answerable enterprise knowledge

[Diagram: Diagram 3 - Enterprise Knowledge AI — diagrams/03-enterprise-knowledge-ai.png]

Enterprise problem

Enterprises have plenty of documents, but employees struggle to find the answers they need quickly. Simple vector search has limits in freshness, permissions, citing sources and handling complex questions.

Latest AI capabilities

- hybrid search: vector + BM25
- Better search quality with rerankers
- metadata/ACL filtering
- citation grounding
- Graph RAG / multi-hop retrieval
- query rewriting and decomposition

Implementation stack

- LlamaIndex, Haystack, LangChain
- Milvus, Qdrant, Weaviate, Chroma
- BGE-M3, E5, Jina embeddings, bge-reranker
- vLLM/SGLang answer generation
- Ragas evaluation

Deep Case Scenario — Knowledge AI for manufacturing equipment failure response

Item	Details
Industry	Manufacturing, semiconductors, equipment operations
Enterprise problem	When equipment fails, field engineers must dig through manuals, past maintenance history and parts documents themselves
Current approach	Reliance on veterans, file server searches, phone inquiries
How AI is applied	Index equipment manuals, error codes, maintenance history and parts catalogs, and answer questions with source citations and step-by-step actions
Tech stack	Docling/PaddleOCR → chunking/metadata → Qdrant/Milvus → BGE-M3 + reranker → LLM answer with citations
GPU workload	Embedding/reranking on L4/L40S/A100, answer generation on L40S/A100/H100, autoscaling for large numbers of concurrent queries
Expected KPIs	Faster failure response, fewer questions to veterans, standardized maintenance procedures, less downtime
gcube.ai point	Integrated document indexing, search, reranking and inference on a single GPU infrastructure

gcube.ai message

gcube.ai turns an enterprise's internal documents not into merely searchable knowledge, but into Enterprise Knowledge AI that leads directly to answers for the job at hand.

4.5 Document Intelligence Pipeline

Converting PDFs, scans, tables and images into AI-ready data

Enterprise problem

Much of enterprise data lives in PDFs, scans, tables, charts and email attachments. When document structuring quality is low, RAG and fine-tuning quality drop too.

Latest AI capabilities

- layout-aware OCR
- table extraction
- key-value extraction
- PDF-to-Markdown/JSON conversion
- VLM-based document and chart understanding
- human validation workflow

Implementation stack

- PaddleOCR, Docling, olmOCR
- Qwen2.5-VL family, LayoutLM/Donut family
- validation rules, schema extraction
- RAG index or ERP/CRM ingestion

Deep Case Scenario — Automated insurance claim document processing

Item	Details
Industry	Insurance, healthcare and financial back offices
Enterprise problem	Claim forms, receipts, medical certificates and test results are entered by hand
Current approach	Partial OCR + human review/entry. Errors rise when tables and scan quality are poor
How AI is applied	Document type classification → OCR/VLM structure extraction → JSON schema validation → human review of anomalies → entry into business systems
Tech stack	PaddleOCR/Docling/olmOCR → Qwen-VL validation → schema parser → human review UI → ERP/insurance system API
GPU workload	A100/H100 for large OCR/VLM batch processing, L40S-class for lightweight OCR, a GPU queue for peak throughput
Expected KPIs	Shorter processing time per document, lower input error rate, higher back-office throughput, better RAG quality
gcube.ai point	Provides a preprocessing GPU pipeline that turns unstructured documents into AI-ready data

gcube.ai message

gcube.ai combines document AI, VLMs, OCR and RAG indexing to turn an enterprise's unstructured documents into AI assets ready for immediate use.

4.6 Generative Content Factory

A generative media engine for mass-producing ad, product and video content

Enterprise problem

Marketing teams must quickly produce ad creatives, thumbnails, product images and short-form videos for each campaign. Outsourced production is costly with long revision lead times, and there aren't enough creatives for A/B testing.

Latest AI capabilities

- node-based generation workflow
- brand LoRA / style consistency
- image-to-video, video-to-video
- Composition, person and product control with ControlNet/IP-Adapter
- batch generation and review queue

Implementation stack

- ComfyUI
- FLUX, Stable Diffusion 3.5, SDXL
- Wan2.1/Wan2.2, CogVideoX, LTX-Video
- ControlNet, IP-Adapter, LoRA
- TTS, lip-sync, audio generation

Deep Case Scenario — E-commerce campaign content factory

Item	Details
Industry	E-commerce and brand marketing
Enterprise problem	With so many products, the team can't produce enough banners, product page images and social short-form videos
Current approach	Designers and agencies produce content manually, focusing on key products
How AI is applied	A ComfyUI workflow generates image candidates from the product DB and campaign brief, and approved images are

	extended into short-form videos
Tech stack	Product DB → prompt template → ComfyUI → FLUX/SD + ControlNet/IP-Adapter → Wan/CogVideoX → review queue
GPU workload	L40S/A100 for image generation, A100/H100 for high-resolution and large batches, H100/H200/B200-class recommended for video generation
Expected KPIs	Lower cost per creative, more creatives per campaign, shorter revision lead time, better CTR/CVR
gcube.ai point	Scales generative media workloads on on-demand GPUs only when needed

gcube.ai message

gcube.ai helps marketing teams shift generative content production from outsourcing to a GPU-Native Content Factory.

4.7 AI QualityOps / LLMOps

An operations layer that measures AI quality, cost and safety

[Diagram: Diagram 5 - AI QualityOps & Data Flywheel — diagrams/05-qualityops-data-flywheel.png]

Enterprise problem

Many teams deploy AI without knowing whether it actually works well. Without continuous measurement of quality, hallucination, cost, latency, user satisfaction and regressions, production AI quickly loses trust.

Latest AI capabilities

- trace-based observability
- RAG evaluation
- LLM-as-judge evaluation
- prompt regression testing
- red teaming
- cost and latency monitoring
- feedback-to-dataset loop

Implementation stack

- Arize Phoenix, Langfuse
- Ragas, Promptfoo, DeepEval, Opik
- OpenTelemetry
- Grafana/Prometheus

Deep Case Scenario — QualityOps for an internal RAG chatbot

Item	Details
Industry	All industries
Enterprise problem	Even when the internal document chatbot gives wrong answers, the operations team can't identify the cause or frequency
Current approach	Manual log checks after user complaints
How AI is applied	Trace every Q&A and automatically evaluate RAG faithfulness/context precision, cost and latency. Failure cases become an improvement backlog
Tech stack	Langfuse/Phoenix trace → Ragas eval → Promptfoo regression → Grafana dashboard → Argilla review
GPU workload	L40S/A100/H100 for evaluation judge models, reranking and batch evaluation
Expected KPIs	Lower hallucination rate, regressions caught in advance, cost optimization, shorter quality improvement cycles
gcube.ai point	Repeatedly runs quality evaluation and retraining loops for production AI services in a GPU environment

gcube.ai message

gcube.ai doesn't stop at deploying AI; it provides an AI QualityOps environment for measuring and improving quality, cost and safety.

4.8 Trust, Safety & AI Governance Layer

Trust-based AI operations for regulated industries

Enterprise problem

Finance, healthcare, public sector, manufacturing and legal organizations must control personal data, access permissions, audit logs, network separation and policy-violating responses. The more AI connects to business systems, the more essential a governance layer becomes.

Latest AI capabilities

- PII detection/masking
- prompt injection/jailbreak defense
- output policy validation
- role-based retrieval filtering
- audit logging
- zero-trust AI access control

Implementation stack

- Microsoft Presidio
- NeMo Guardrails, Guardrails AI, LLM Guard
- OPA(Open Policy Agent)
- Llama Guard family
- private deployment / on-premise integration

Deep Case Scenario — Patient data protection AI for a healthcare provider

Item	Details
Industry	Healthcare and hospitals
Enterprise problem	They want to use documents containing patient information with AI, but privacy and internal network requirements are strict
Current approach	Sensitive documents are excluded from AI adoption, or de-identified manually
How AI is applied	Detect and mask PII at input, allow document search only for authorized users, and store every response and access record in audit logs
Tech stack	SSO/RBAC → Presidio → OPA policy → RAG ACL filter → NeMo Guardrails → audit log
GPU workload	Runs RAG over sensitive documents and medical-domain LLM/VLM inference on dedicated GPUs
Expected KPIs	Lower risk of exposing sensitive information, faster audit response, broader use of AI on the internal network
gcube.ai point	Lowers the barriers to AI adoption in regulated industries with a GPU runtime suited to private/on-premise use

gcube.ai message

gcube.ai lets enterprises use the latest AI capabilities in an internally controlled environment without giving up security or compliance.

4.9 Synthetic Data Flywheel

A continuous improvement loop that turns failure data into training assets

Enterprise problem

Enterprises want better AI quality but lack high-quality training data. Real customer data is sensitive, and domain experts don't have time for labeling. So a loop is needed that turns operational logs and failure cases into data assets.

Latest AI capabilities

- synthetic instruction generation
- AI feedback and human review
- preference dataset construction
- DPO/ORPO alignment
- retrieval/reranker training data generation
- eval failure mining

Implementation stack

- Argilla, distilabel
- Failure logs from Phoenix/Langfuse/Ragas
- Unsloth/Axolotl/LLaMA Factory
- Promptfoo regression suite

Deep Case Scenario — Data flywheel for legal contract review

Item	Details
Industry	Legal, professional services, corporate legal teams
Enterprise problem	They want to build contract review AI but lack clause-level risk labels and high-quality revision data
Current approach	Lawyers review contracts manually, and review comments are scattered across documents
How AI is applied	Anonymize existing review comments, have AI generate clause-level risk questions, answers and revision candidates, and accumulate lawyer-reviewed results as a preference dataset
Tech stack	Presidio de-identification → distilabel synthetic data → Argilla human review → DPO/LoRA tuning → Promptfoo eval
GPU workload	A100/H100-class GPUs for synthetic generation, judge evaluation and fine-tuning

Expected KPIs	Lower cost of building training data, shorter time to first-draft contract reviews, higher detection rate for recurring risks
gcube.ai point	Runs data generation, review, training and evaluation as a single GPU-based flywheel

gcube.ai message

gcube.ai turns failures and feedback from production back into training data, building enterprise AI that gets smarter over time.

4.10 GPU-Native AI Operating Platform

Scaling PoCs into a company-wide AI operating platform

Enterprise problem

Departmental AI PoCs keep growing, but every operating environment is different. If model versions, GPU usage, permissions, evaluation results and cost policies aren't unified, AI remains a collection of experiments rather than a company-wide platform.

Latest AI capabilities

- Kubernetes-based GPU orchestration
- autoscaling and quota
- model registry and deployment pipeline
- multi-model serving
- workload-aware GPU scheduling
- observability and chargeback

Implementation stack

- Kubernetes + NVIDIA GPU Operator
- KServe, BentoML, Ray Serve
- Triton Inference Server
- MLflow/Kubeflow
- Prometheus/Grafana/OpenTelemetry
- LiteLLM/API Gateway
- Dify/Open WebUI/AnythingLLM

Deep Case Scenario — Company-wide AI Portal & Runtime

Item	Details
Industry	Large enterprises, public sector, group IT
Enterprise problem	Each department builds its own AI PoC, duplicating models, data, GPUs and security policies
Current approach	VM/GPU allocation per project, manual deployment, hard-to-track costs
How AI is applied	Provide a common AI portal and API gateway so each department requests, deploys and monitors RAG/agent/LLM APIs in a standard way
Tech stack	AI Portal → API Gateway → KServe/Ray Serve/vLLM/SGLang → MLflow Registry → QualityOps/Governance
GPU workload	GPU scheduling by workload: daytime inference, overnight training, batch evaluation, media generation and more
Expected KPIs	Higher PoC-to-production conversion, shorter model deployment lead time, higher GPU utilization, cost visibility by department
gcube.ai point	Positioned not as single-GPU rental but as the runtime layer of a company-wide AI operating platform

gcube.ai message

gcube.ai provides the GPU-Native foundation for enterprises to unify multiple AI PoCs into a single Enterprise AI Operating Platform.

5. Recommended workloads by gcube.ai GPU lineup

Note: The tables below are a recommended architecture matrix for this whitepaper draft. The actual GPU lineup, VRAM, networking, hourly pricing and available regions offered by gcube.ai must be confirmed before publication.

5.1 GPU class definitions

GPU Class	Example GPUs	General characteristics	Best suited for
Entry / Dev GPU	RTX 4090-class, L4-class	Development, prototyping, lightweight inference	PoC, embedding, small model test
Inference GPU	L40S-class,	Cost-efficient	7B~14B LLM, RAG,

	A10/A30-class	inference, image generation	ComfyUI
Training / Balanced GPU	A100 80GB-class	Balanced training and inference, large VRAM	LoRA, 32B~70B serving, VLM
High-Performance GPU	H100/H200-class	High-performance training and inference, FP8 optimization	Large-scale LLMs, high-throughput serving, video
Frontier GPU	B200/GB200-class	Blackwell generation, ultra-high-performance inference and training	Large-scale enterprise AI platforms, frontier workloads

5.2 Recommended GPUs by workload

Workload	Recommended GPU class	VRAM guideline	Recommended stack	Example use	Design point
RAG embedding	Entry/ Inference	16~48GB+	BGE-M3, E5, Jina, TEI	Indexing internal documents	Optimize batch processing and vector DB I/O
RAG reranking	Inference/ Balanced	24~80GB+	bge-reranker, cross-encoder	Improving search accuracy	Balance latency and quality
7B~14B LLM serving	Inference/ Balanced	24~80GB+	vLLM, SGLang	Internal chatbots, summarization	continuous batching, quantization
32B~70B LLM serving	Balanced/ High-Performance	80GB+ multi-GPU	vLLM TP, SGLang, TensorRT-LLM	High-quality QA, legal/medical	tensor parallel, KV cache
100B+ large model serving	High-Performance/ Frontier	multi-GPU	TensorRT-LLM, vLLM cluster	Advanced reasoning	network fabric, sharding
LoRA/QLoRA fine-tuning	Balanced/ High-Performance	48~80GB+	Unsloth, LLaMA Factory, Axolotl	Domain LLMs	Prioritize data quality and evals
Full fine-tuning	High-Performance/ Frontier	multi-node	Axolotl, DeepSpeed, FSDP	Large dedicated models	Rigorously review cost and data scale
Document AI/OCR	Inference/ Balanced	24~80GB+	PaddleOCR, Docling, Qwen-VL	Contracts, claim forms	batch queue, schema validation
Image generation	Inference/ Balanced	24~80GB+	ComfyUI, FLUX, SD3.5	Ad creatives	workflow template,

					review queue
Video generation	High-Performance/ Frontier	80GB+	Wan, CogVideoX, LTX-Video	Short-form, virtual humans	Long processing times and batch scheduling
AI Agent Runtime	Inference/ High-Performance	Varies by scale	LangGraph, SGLang, vLLM	Research/ sales/coding agents	tool latency, trace, approval
AI QualityOps evaluation	Inference/ Balanced	24~80GB+	Phoenix, Ragas, Promptfoo	batch eval, judge model	Overnight evaluation based on production logs
Enterprise AI Platform	Balanced~Fr ontier	mixed pool	KServe, Ray Serve, MLflow, Gateway	Company-wide AI portal	Separate GPU pools and quotas

5.3 Quick matrix for purchase and consultation

Customer situation	Recommended starting point	Expansion direction
"We want to start with an internal document chatbot"	RAG + 7B/14B serving on Inference GPU	Add a reranker, QualityOps and ACL filtering
"Our external API costs are too high"	Private AI Gateway + vLLM/SGLang	Model routing, caching, per-department quotas
"We want our support bot to match our company's style"	LoRA fine-tuning on A100/H100	DPO, synthetic data flywheel
"We want to mass-produce ad images and short-form videos"	ComfyUI on L40S/A100	H100/H200/B200 pool for video generation
"We're in finance/healthcare, so security comes first"	Private deployment + Governance Layer	Presidio, OPA, Guardrails, audit log
"We want to bring several departmental PoCs together into one operating platform"	Enterprise AI Gateway + ModelOps	GPU pool, MLflow, KServe/Ray Serve, QualityOps

6. Reference Architectures on gcube.ai

6.1 Private LLM Serving Architecture

[Diagram: Private AI Gateway — diagrams/02-private-ai-gateway.png]

The key is that every AI request passes through the gateway, leaving a record of authentication,

policy, masking, model routing, cost and traces. This lets enterprises reduce external API dependence and offer an internal standard AI API.

6.2 Enterprise RAG Architecture

[Diagram: Enterprise Knowledge AI — diagrams/03-enterprise-knowledge-ai.png]

RAG must include more than simple vector search: document structuring, permission filtering, hybrid retrieval, reranking, citation grounding and evaluation.

6.3 Agentic Workflow Architecture

[Diagram: Agentic Workflow Engine — diagrams/04-agentic-workflow-engine.png]

For business agents, controllable autonomy matters more than full autonomy. High-risk actions go through human approval, and every tool call must be auditable.

6.4 QualityOps & Data Flywheel Architecture

[Diagram: AI QualityOps Data Flywheel — diagrams/05-qualityops-data-flywheel.png]

Operational logs are both cost data and quality data, and failure cases become the next training data. The more this loop accumulates, the more enterprise AI evolves to fit the domain.

6.5 Sovereign AI Stack Architecture

[Diagram: Sovereign AI Stack — diagrams/01-sovereign-ai-stack.png]

The Sovereign AI Stack connects models, knowledge, agents, quality, security and the GPU runtime into a single layered operating system.

7. Industry Playbooks

Finance

- Control external API use with a Private AI Gateway
- RAG-based QA on product terms and regulations
- PII masking and audit logs

- Support quality evaluation and regression testing

Healthcare

- De-identification of patient information
- Document Intelligence for test result reports
- RAG based on medical guidelines
- Internal network / private deployment

Manufacturing

- Knowledge AI for equipment manuals
- OCR/VLM processing of inspection reports
- Failure response agents
- Multimodal Q&A for field workers

Public sector & education

- AI search for civil petition documents
- Internal knowledge portal on a secure network
- Educational content generation
- Policy-compliant AI response management

E-commerce & marketing

- Generative Content Factory
- Automated generation of product descriptions, banners and short-form videos
- Customer VOC analysis
- Domain tuning for brand tone

Legal & professional services

- Contract risk review AI
- Knowledge AI for case law and contracts
- Document summaries and revision suggestions
- Data flywheel built on expert review

8. ROI Measurement Framework

8.1 Cost savings

Monthly cost savings = external API cost savings + outsourcing cost savings + manual labor cost savings – gcube operating costs

8.2 Time savings

Time saved = original processing time – processing time after AI adoption
Monetary value = time saved × staff hourly cost × monthly volume

8.3 Productivity gains

Productivity gain rate = throughput after AI adoption / original throughput – 1

8.4 Risk reduction

- Blocking sensitive information from leaving the company
- Faster audit response
- Blocking non-compliant responses
- Catching regressions in advance when models change

8.5 New revenue opportunities

- Faster customer response
- Personalized content and recommendations
- Launching new AI-based services
- Shorter product launch lead times

9. Adoption Roadmap

Phase	Goal	Duration	Key deliverables
Phase 1. AI Workload Discovery	Identify high-ROI task candidates	1–2 weeks	use case list, data readiness
Phase 2. Private AI PoC	Validate value on a single task	2–6 weeks	PoC, baseline KPI
Phase 3. Department Pilot	Validate operations at the department level	1–3 months	limited production, logs, eval

Phase 4. Production AI Runtime	Production with SLAs and security	3–6 months	Gateway, monitoring, governance
Phase 5. Enterprise AI Operating Platform	Company-wide AI platform	6 months+	AI Portal, ModelOps, GPU pool, QualityOps

10. Technical Appendix

10.1 Implementation stack candidates

Area	Enterprise term	Implementation candidates
Model Serving	Private AI Gateway	vLLM, SGLang, TensorRT-LLM, TGI, LiteLLM
Knowledge AI	Enterprise Knowledge Layer	LlamaIndex, Haystack, Milvus, Qdrant, Weaviate
Agent	Agentic Workflow Engine	LangGraph, CrewAI, AutoGen, MCP
Document AI	Document Intelligence Pipeline	PaddleOCR, Docling, olmOCR, Qwen-VL
Content AI	Generative Content Factory	ComfyUI, FLUX, SD3.5, Wan, CogVideoX
QualityOps	AI Quality Control Layer	Phoenix, Langfuse, Ragas, Promptfoo
Governance	Trust & Governance Layer	Presidio, NeMo Guardrails, OPA, LLM Guard
Data Flywheel	Synthetic Data Flywheel	Argilla, distilabel, DPO/ORPO pipeline
Platform	AI Operating Platform	Kubernetes, KServe, Ray Serve, BentoML, MLflow

10.2 Items to verify before publication

- gcube.ai's actual GPU lineup, VRAM, pricing, networking and storage options
- Latest licenses for each model and tool
- Latest benchmark figures: vLLM/SGLang/TensorRT-LLM, FLUX/Wan/CogVideoX, etc.
- Whether customer cases can be cited
- Legal review of security claims for finance, healthcare and the public sector

11. Final Message

The essence of enterprise AI in 2026 is not “we adopted AI” but being able to run AI. Enterprises must move beyond choosing models and bring models, knowledge, agents, quality, security and GPU infrastructure together into a single execution system.

The Sovereign AI Stack is a strategy for enterprises to regain control over how their AI runs. And gcube.ai is the GPU-Native AI Runtime for putting this strategy into practice in real business environments.

Move from an enterprise that calls external APIs to an enterprise that runs its own AI services.