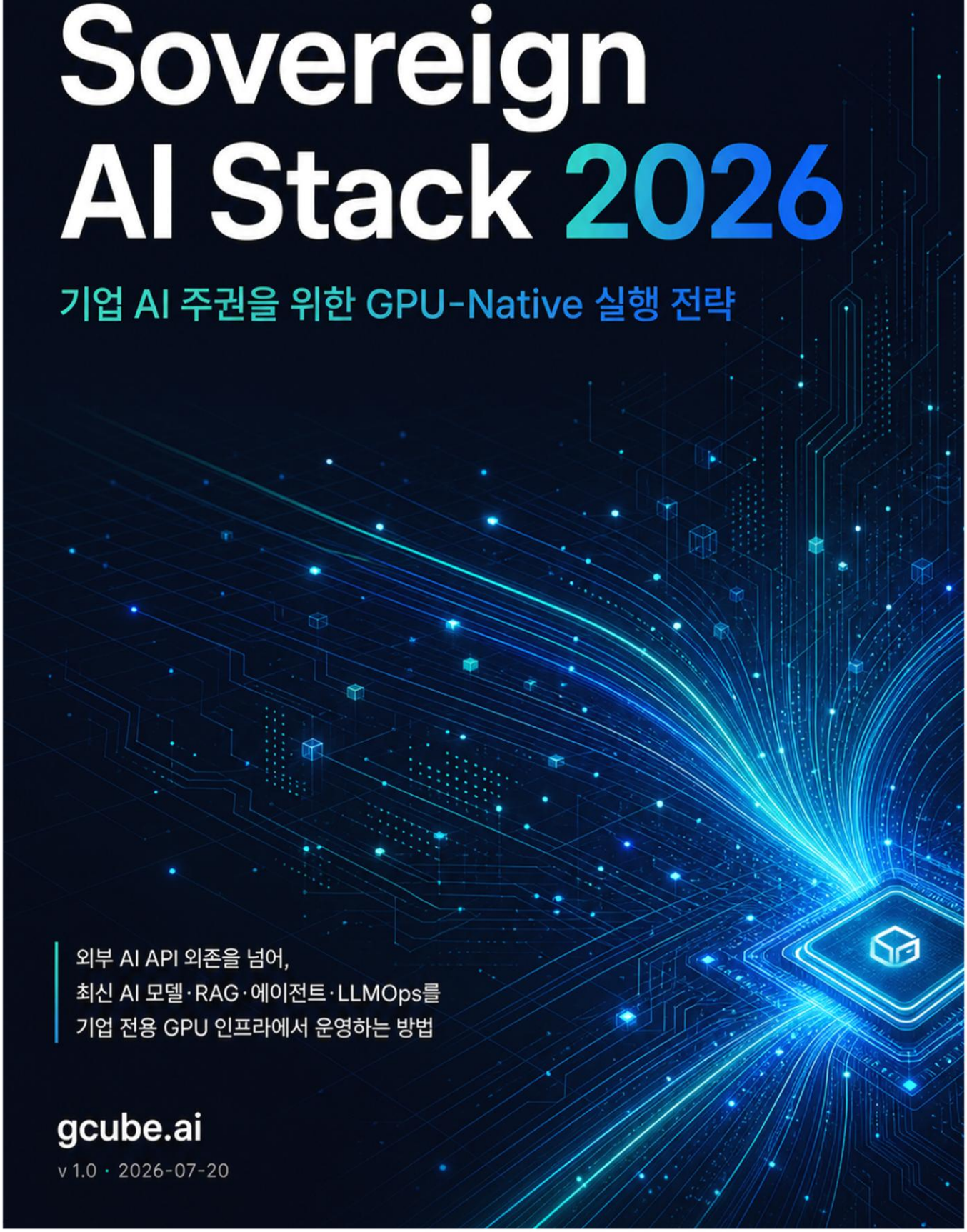
GCUBE.AI WHITEPAPER · 2026

Sovereign AI Stack 2026

기업 AI 주권을 위한 GPU-Native 실행 전략



외부 AI API 외존을 넘어,
최신 AI 모델·RAG·에이전트·LLMOps를
기업 전용 GPU 인프라에서 운영하는 방법

gcube.ai

v 1.0 · 2026-07-20

GCUBE.AI WHITEPAPER · 2026

Sovereign AI Stack 2026

기업 AI 주권을 위한 GPU-Native 실행 전략

외부 AI API 의존을 넘어, 최신 AI 모델·RAG·에이전트·LLMOps 를 기업 전용 GPU 인프라에서 운영하는 방법

발행 주체	gcube.ai
문서 버전	v 1.0
작성일	2026-07-20
대상 독자	CIO · CTO · CDO · AI Transformation 리더 · 보안/인프라 책임자 · 현업 자동화 리더

"외부 API 를 호출하는 기업에서, 자체 AI 서비스를 운영하는 기업으로 전환하십시오."

2026 년 기업 AI 경쟁력은 어떤 모델을 쓰느냐가 아니라, AI 를 얼마나 안전하고 효율적으로 자신의 데이터·보안·인프라 안에서 운영할 수 있느냐에 달려 있습니다.

핵심 포지셔닝

gcube.ai 는 기업이 최신 AI 모델과 에이전트, RAG, LLMOps 를 자체 데이터·보안·GPU 인프라 안에서 운영할 수 있게 하는 GPU-Native AI Runtime 입니다.

※ 본 문서에 수록된 Case Study 는 공개된 산업별 대표 사례 및 시장 데이터를 바탕으로 재구성된 참조용(Reference) 시나리오입니다. 수치는 동종 업계 유사 프로젝트에서 일반적으로 관측되는 개선 범위를 나타냅니다.

SECTION 01

Executive Brief

생성형 AI 의 첫 번째 물결은 외부 API 와 SaaS 도구를 통해 시작되었다. 많은 기업은 ChatGPT, Claude, Copilot, 이미지 생성 도구, AI 검색 서비스를 빠르게 도입하며 생산성 가능성을 확인했다. 하지만 PoC 가 실제 운영으로 확장되는 순간 기업은 새로운 병목을 만난다.

- 외부 API 토큰 비용이 사용량과 함께 증가한다.
- 고객 데이터, 내부 문서, 소스코드, 의료·금융 정보가 외부로 전송된다.
- 범용 모델은 기업 내부 용어, 정책, 업무 맥락을 충분히 이해하지 못한다.
- AI 답변 품질, 할루시네이션, 지연시간, 비용을 지속적으로 측정하지 못한다.
- 부서별 PoC 가 늘어날수록 모델, 데이터, 권한, GPU 자원이 파편화된다.

따라서 2026 년 기업 AI 의 핵심 전환은 AI API 소비에서 Private AI Runtime 운영으로 이동하는 것이다. 기업은 최신 AI 모델을 단순히 호출하는 것이 아니라, 자신의 데이터·권한·보안·평가 체계 안에서 운영해야 한다. 이 백서에서 말하는 Sovereign AI Stack 은기업이 AI 실행권을 되찾기 위한 전체 운영 체계다.

Sovereign AI Stack =

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

gcube.ai 는 기업이 최신 AI 스펙을 실험 단계에서 끝내지 않고, 실제 업무 생산성·비용 절감·보안 강화·신규 서비스 창출로 연결할 수 있도록 지원하는 GPU-Native AI Runtime 이다.

SECTION 02

The Enterprise AI Shift

2026 년 기업 AI 의 5 대 병목

병목	기존 증상	Sovereign AI Stack 해결 방향
데이터 보안	민감 데이터를 외부 API 에 넣기 어려움	Private AI Gateway · private deployment · PII masking
비용 예측성	토큰 과금 증가, SaaS 라이선스 중복	GPU-Native Runtime · model routing · caching · quota
도메인 정확도	내부 용어·정책·문서 맥락 부족	Domain-Adaptive Intelligence · Knowledge AI
운영 품질	배포 후 품질·환각·지연시간 측정 불가	AI QualityOps · tracing · eval · regression test
확장성	부서별 PoC 가 플랫폼으로 연결되지 않음	AI Operating Platform · ModelOps · governance

SECTION 03

Sovereign AI Stack Framework

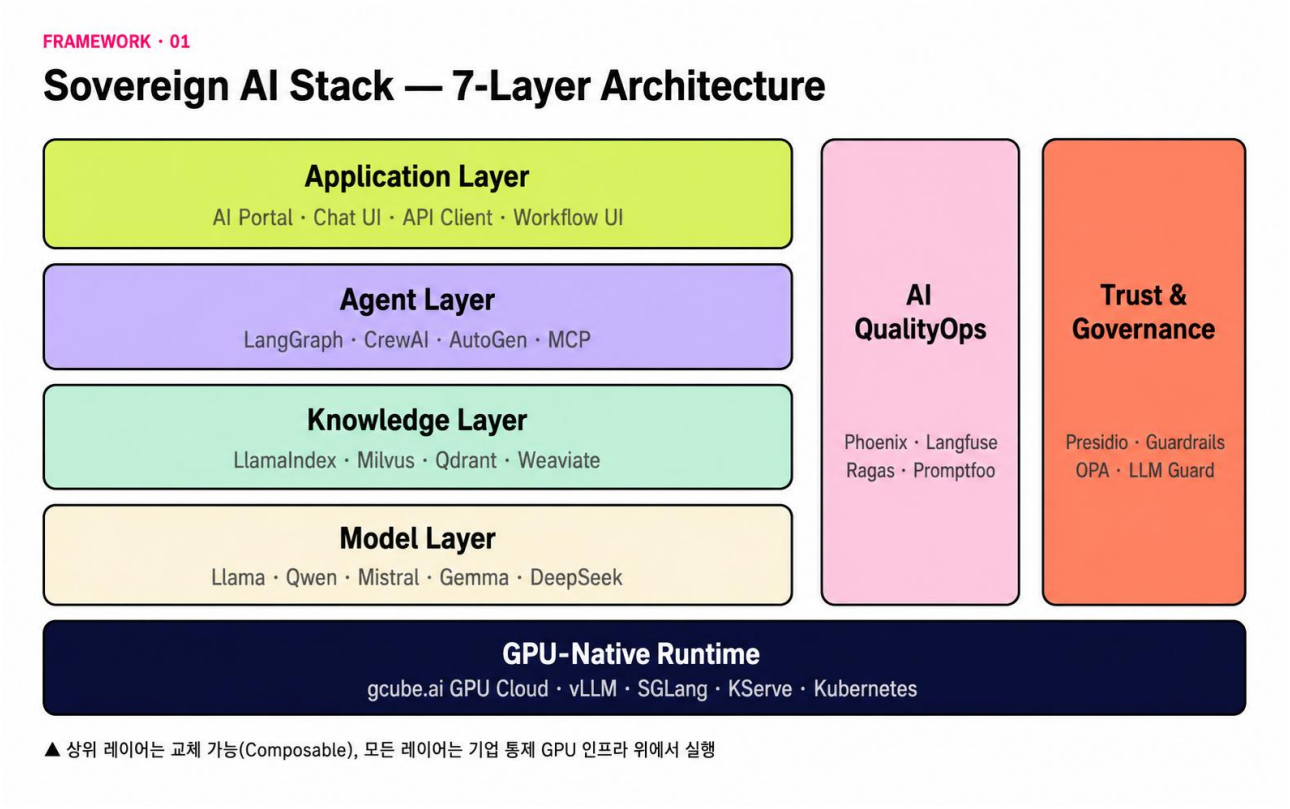


Diagram 1 — Sovereign AI Stack 7-Layer Architecture

3.1 — 7-Layer Framework

각 대표 레이어의 기술은 다음과 같다.

Layer	역할	대표 기술
Application	직원·고객·업무 앱이 AI 를 사용하는 접점	AI Portal · Chat UI · API Client
Agent	목표를 계획하고 도구를 호출하는 실행 계층	LangGraph · CrewAI · AutoGen · MCP
Knowledge	사내 문서·데이터를 검색 가능한 지식으로 전환	LlamaIndex · Milvus · Qdrant · Weaviate
Model	기업 업무를 이해하고 생성·분석하는 모델 계층	Llama · Qwen · Mistral · Gemma · DeepSeek
QualityOps	품질·비용·지연시간·회귀 테스트 운영	Phoenix · Langfuse · Ragas · Promptfoo
Trust & Governance	개인정보·권한·감사·안전 정책	Presidio · Guardrails · OPA · LLM Guard
GPU-Native Runtime	학습·추론·문서·비디오·에이전트 실행	gcube.ai GPU Cloud · vLLM · SGLang · KServe

3.2 — 핵심 원칙

원칙	내용
1. Private by Design	민감 데이터는 가능한 한 기업 통제 환경 안에서 처리한다.
2. Composable by Architecture	모델·RAG·에이전트·평가·보안을 교체 가능한 레이어로 설계한다.
3. GPU-Native by Operation	AI 워크로드별 GPU 특성과 비용 구조를 기준으로 운영한다.
4. QualityOps by Default	배포 후 품질·비용·안전성을 지속 측정한다.
5. Human-Governed Autonomy	에이전트는 자동 실행하되, 고위험 액션은 사람 승인과 감사 로그를 남긴다.

SECTION 04

10 Strategic AI Workloads — Anonymized Case Studies

각 워크로드는 [기업 문제 → 핵심 스펙 → 익명 고객 Case Study] 순으로 구성된다. Case Study 는 고객 비밀유지를 위해 익명 처리된 산업 대표 시나리오이며, 수치는 동일 유형 프로젝트의 일반적 개선 범위를 반영한다.

4.1 Domain-Adaptive Intelligence

기업 데이터를 이해하는 전용 언어지능

기업 문제

범용 LLM 은 일반 지식에는 강하지만 기업 내부 용어, 상품명, 상담 정책, 법무·의료·제조 맥락을 정확히 반영하지 못한다. 프롬프트만으로는 답변 스타일, 규정 준수, 전문 용어 사용을 안정적으로 통제하기 어렵다.

구분	내용
핵심 AI 스펙	LoRA/QLoRA 기반 PEFT · DPO/ORPO 선호도 정렬 · instruction/preference dataset 분리 운영 · 한국어 도메인 튜닝 · 평가셋 기반 regression test 후 배포
구현 스택	Unsloth · LLaMA Factory · Axolotl · PEFT/TRL · bitsandbytes · FlashAttention · vLLM/SGLang 서빙

CASE STUDY — ANONYMIZED

국내 대형 통신사 T 사 — 고객센터 전용 상담 LLM

산업	규모	도입 범위	GPU 구성
통신 (국내 이동통신 상위권)	상담사 약 1,800 명 · 월 상담 120 만 건	상담 답변 초안 생성 · 정책 준수 검증	학습 A100 80GB×4 · 운영 추론 L40S 풀

도입 배경 · 과제

T 사 고객센터는 요금제 개편·결합상품·위약금 정책이 분기마다 바뀌는 환경에서 상담사마다 답변 품질 편차가 컸다. 외부 API 기반 챗봇을 시범 도입했지만 자사 요금제 명칭을 혼동하고 금지 표현(확정적 보상 약속 등)을 걸러내지 못했으며, 상담 로그를 외부로 전송하는 것 자체가 보안 심의를 통과하지 못했다.

GPU-Native 실행 전략

우수 상담 로그 12 만 건과 FAQ·정책 문서를 Presidio 로 비식별화한 뒤 Argilla 에서 품질 검수를 거쳐 instruction/preference dataset 으로 변환했다. Unsloth 기반 LoRA/DPO 튜닝으로 자사 정책·금지 표현·상담 톤을 학습시키고, Promptfoo 회귀 테스트를 통과한 버전만 vLLM 으로 배포했다. 전 과정이 gcube.ai 전용 GPU 환경 내부에서 수행되어 상담 데이터는 한 번도 외부로 나가지 않았다.

PIPELINE Presidio 비식별화 → Argilla 검수 → Unsloth LoRA/DPO → Promptfoo 평가 → vLLM 서빙

도입 성과

지표	도입 전	도입 후
평균 답변 작성 시간	상담 건당 약 4.5 분	약 2.4 분 (47% 단축)
재문의율	18%	11%
금지 표현 노출	월 수십 건 수기 적발	배포 전 자동 차단
외부 API 비용	사용량 비례 증가	전용 GPU 정액 운영

Key Point — 민감 상담 데이터의 외부 전송 없이 학습·평가·서빙을 단일 전용 GPU 환경에서 통합 운영. 분기별 정책 개편 시 재튜닝 주기를 2 주 이내로 단축.

"gcube.ai 는 기업이 자신의 데이터로 전용 언어지능을 만들고, 외부 API 가 아닌 Private AI Runtime 에서 안전하게 운영할 수 있도록 최소 비용으로 효율적인 GPU 서비스를 제공합니다."

4.2 Private AI Gateway

외부 API 의존을 대체하는 사내 AI API 계층

기업 문제

외부 LLM API 는 빠르게 시작할 수 있지만, 사용량 증가와 함께 비용·보안·SLA 문제가 커진다. 기업은 다양한 모델을 하나의 API 로 통합하고 권한·비용·로그·품질을 중앙에서 관리해야 한다.

구분	내용
핵심 AI 스펙	OpenAI-compatible API · continuous batching · prefix caching · speculative decoding · FP8/NVFP4 양자화 · model routing/fallback · quota/chargeback/tenant isolation
구현 스택	vLLM · SGLang · TensorRT-LLM · LiteLLM/API Gateway · Prometheus/Grafana · OpenTelemetry · Phoenix/Langfuse

CASE STUDY — ANONYMIZED

국내 상위권 증권사 S 사 — 전사 표준 AI Gateway

산업	규모	도입 범위	GPU 구성
금융 (증권·자산관리)	임직원 약 3,200 명 · 전 부서 대상	전사 LLM API 단일화 · 비용/보안 통제	L40S 추론 풀 + A100/H100 TP 클러스터

도입 배경 · 과제

S 사는 리서치·컴플라이언스·IT 부서가 각자 외부 LLM SaaS 계정을 사용하면서 고객 계좌 정보가 프롬프트에 입력되는 사고 위험과 월 단위로 불어나는 구독·토큰 비용을 동시에 안게 됐다. 금융감독 규제상 데이터 이동 경로를 소명해야 했지만 부서별 SaaS 사용 내역은 추적이 불가능했다.

GPU-Native 실행 전략

모든 LLM 호출을 사내 표준 Gateway 하나로 수렴시켰다. 요청은 SSO/RBAC 인증과 Presidio PII 마스킹을 거쳐 업무 유형별로 7B~70B 사내 모델에 라우팅되고, 부서별 쿼터·비용이 자동 정산된다. 모든 요청·응답은 Langfuse trace 와 감사 로그로 남아 규제 대응 자료로 즉시 활용된다.

PIPELINE API Gateway/LiteLLM → Presidio → vLLM/SGLang 모델 풀 → Phoenix/Langfuse → Grafana

도입 성과

지표	도입 전	도입 후
외부 LLM API 호출	전 부서 무통제 사용	78% 감소 (잔여분은 승인제)
민감정보 외부 전송	추적 불가	0 건 (게이트웨이 차단)
부서별 AI 비용	합산 불가	월 단위 chargeback 리포트
응답 지연 (p95)	외부 API 변동	2.1 초 내 SLA 관리

Key Point — 전용 GPU 풀과 API 계층을 결합해 '금지'가 아닌 '안전한 표준 경로'를 제공. 도입 3 개월 만에 임직원 60% 이상이 주 1 회 이상 사용하는 전사 인프라로 정착.

"외부 AI API 를 통제 불가능하게 호출하는 구조에서 벗어나, gcube.ai 위에 기업 전용 Private AI Gateway 를 구축 및 운영할 수 있습니다."

4.3 Agentic Workflow Engine

판단하고 실행하는 자율 업무 엔진

기업 문제

RPA 는 정해진 절차에는 강하지만 예외 상황, 자연어 판단, 문서 이해, 도구 선택에는 약하다. 기업은 검색·작성·검토·시스템 입력·승인 요청을 하나의 업무 단위로 자동화할 필요가 있다.

구분	내용
핵심 AI 스펙	stateful agent graph · function calling/tool use · MCP 도구 연결 · multi-agent orchestration · human-in-the-loop approval · traceable action log
구현 스택	LangGraph · CrewAI · AutoGen · MCP · Playwright · SGLang/vLLM tool-use 서빙 · Phoenix/Langfuse agent trace

CASE STUDY — ANONYMIZED

글로벌 산업설비 제조사 M 사 한국법인 — B2B 제안서 에이전트

산업	규모	도입 범위	GPU 구성
B2B 제조 (산업설비·플랜트)	영업조직 약 60 명 · 연 제안 800 여 건	제안서 초안 · 고객 리서치 · 후속 메일	L40S/A100 autoscaling 추론 풀

도입 배경 · 과제

M 사 영업 담당자는 제안서 한 건을 만들기 위해 CRM 이력 조회, 고객사 산업 리서치, 과거 유사 프로젝트 검색, 견적 템플릿 작성을 오가며 평균 3 일을 소모했다. 성수기에는 제안 요청을 소화하지 못해 기회를 흘려보냈고, 담당자별 제안서 품질 편차도 컸다.

GPU-Native 실행 전략

LangGraph 기반 에이전트가 CRM API 로 고객 이력을 조회하고, 산업별 pain point 를 리서치한 뒤, 사내 수주 제안서·성공 사례를 Enterprise Knowledge AI 로 불러와 표준 템플릿에 맞는 초안을 작성한다. 초안과 후속 메일은 반드시 담당자 승인을 거쳐 발송되며, 에이전트의 모든 도구 호출은 감사 로그로 기록된다.

PIPELINE LangGraph planner → CRM API → Knowledge AI(RAG) → 초안 생성 → human approval → audit log

도입 성과

지표	도입 전	도입 후
제안서 초안 소요	평균 3 일	약 4 시간 (85% 단축)
월 제안 처리량	1 인당 1.1 건	1 인당 1.9 건 (1.7 배)
후속 메일 리드타임	평균 2.5 일	당일 발송
제안서 품질 편차	담당자별 상이	표준 템플릿 준수율 96%

Key Point — 에이전트 실행 중 발생하는 다수의 짧은 추론·검색 워크로드를 autoscaling GPU 풀에서 처리해 동시 사용자 증가에도 응답 안정성 유지. 고위험 액션은 전량 사람 승인.

"gcube.ai 는 AI 에이전트가 업무 시스템과 연결되어 판단·작성·실행을 수행하는 데 필요한 GPU-Native 실행 환경을 제공합니다."

4.4 Enterprise Knowledge AI

사내 문서를 답변 가능한 기업 지식으로 전환

기업 문제

기업 문서는 많지만 직원이 원하는 답을 빠르게 찾기 어렵다. 단순 벡터 검색은 최신성, 권한, 근거 제시, 복합 질문 처리에서 한계가 있다.

구분	내용
핵심 AI 스펙	hybrid search (vector+BM25) · reranker · metadata/ACL filtering · citation grounding · Graph RAG/multi-hop · query rewriting
구현 스택	LlamaIndex · Haystack · Milvus/Qdrant · BGE-M3 · bge-reranker · vLLM/SGLang · Ragas 평가

CASE STUDY — ANONYMIZED

반도체 장비 운영사 K 사 — 장애 대응 Knowledge AI

산업	규모	도입 범위	GPU 구성
제조 (반도체 장비 운영)	관리 장비 약 2,400 대 · 현장 엔지니어 500 여 명	장비 매뉴얼·정비 이력 질의응답	임베딩/리랭킹 L40S · 답변 생성 A100

도입 배경 · 과제

K 사는 장비 장애가 발생하면 엔지니어가 수천 페이지 매뉴얼과 파일 서버의 정비 이력을 뒤지거나 숙련자에게 전화로 물어야 했다. 야간·주말 장애는 대응이 늦어져 라인 다운타임으로 직결됐고, 숙련 엔지니어의 퇴직과 함께 노하우가 유실되고 있었다.

GPU-Native 실행 전략

장비 매뉴얼, 장애 코드, 10 년치 정비 이력, 부품 카탈로그를 Docling·OCR 로 구조화해 인덱싱했다. 현장 엔지니어가 자연어로 질문하면 hybrid search 와 reranker 가 관련 문서를 찾고, LLM 이 원문 근거 인용과 함께 조치 순서를 제시한다. 장비 담당 권한에 따라 접근 가능한 문서가 ACL 로 필터링되며, 전체 인프라는 사내 보안망 내 전용 GPU 에서 운영된다.

PIPELINE Docling/PaddleOCR → chunking/metadata → Qdrant → BGE-M3 + reranker → 근거 인용 답변

도입 성과

지표	도입 전	도입 후
평균 장애 대응(MTTR)	4.2 시간	2.6 시간 (38% 단축)
숙련자 에스컬레이션	장애의 60%	33% (45% 감소)
야간 장애 초동 대응	익일 오전	10 분 내 조치 가이드
정비 절차 표준화	개인 노하우 의존	근거 문서 기반 표준 절차

Key Point — 문서 인덱싱·검색·리랭킹·추론을 하나의 GPU 인프라에서 통합 운영. 퇴직으로 유실되던 숙련 노하우가 검색 가능한 기업 자산으로 축적.

4.5 Document Intelligence Pipeline

PDF·스캔·표·이미지를 AI-ready 데이터로 변환

기업 문제

기업 데이터의 상당 부분은 PDF, 스캔본, 표, 차트, 이메일 첨부파일로 존재한다. 문서 구조화 품질이 낮으면 RAG 와 파인튜닝 품질도 낮아진다.

구분	내용
핵심 AI 스펙	layout-aware OCR · table extraction · key-value extraction · PDF-to-Markdown/JSON · VLM 문서 이해 · human validation workflow
구현 스택	PaddleOCR · Docling · olmOCR · Qwen2.5-VL · LayoutLM/Donut · schema validation · ERP/CRM ingestion

CASE STUDY — ANONYMIZED

대형 생명보험사 L 사 — 보험금 청구 문서 자동 처리

산업	규모	도입 범위	GPU 구성
보험 (생명보험 상위권)	청구 문서 월 85 만 건 · 심사 인력 300 여 명	청구서·진단서·영수증 구조화 입력	batch OCR/VLM A100 · 경량 처리 L40S

도입 배경 · 과제

L 사 보험금 심사팀은 청구서, 진단서, 진료비 영수증, 검사 결과지를 사람이 육안으로 확인해 시스템에 수기 입력했다. 스캔 품질이 낮은 문서와 병원별 상이한 서식 때문에 입력 오류가 반복됐고, 청구 급증 시기에는 지급 지연 민원이 급증했다.

GPU-Native 실행 전략

접수 문서를 유형별로 자동 분류한 뒤 OCR 과 VLM 으로 항목·표를 구조화하고 JSON schema 검증을 통과한 데이터만 심사 시스템에 자동 입력한다. 신뢰도가 낮은 건만 human review 큐로 분기해 사람은 예외 처리에 집중한다.
진단명·주민번호 등 민감 정보는 전 과정이 내부 GPU 환경에서 처리된다.

PIPELINE 문서 분류 → PaddleOCR/Docling → Qwen-VL 검증 → schema parser → human review → 심사 시스템 API

도입 성과

지표	도입 전	도입 후
문서 1 건 처리 시간	평균 12 분 (수기)	90 초 (자동+예외검수)
입력 오류율	4.8%	0.9%
심사팀 처리량	기준 1.0	3.2 배
지급 지연 민원	성수기 급증	62% 감소

Key Point — 대량 문서 batch 처리 시기에만 GPU 큐를 확장하는 온디맨드 운영으로 피크 비용 최적화. 구조화된 데이터는 RAG·튜닝 학습 자산으로 재활용.

"gcube.ai 는 문서 AI, VLM, OCR, RAG 인덱싱을 결합해 기업의 비정형 문서를 즉시 활용 가능한 AI 자산으로 전환할 수 있는 GPU 클라우드 서비스를 제공합니다"

4.6 Generative Content Factory

광고·상품·영상 콘텐츠를 대량 생산하는 생성형 미디어 엔진

기업 문제

마케팅 조직은 캠페인별 광고 소재, 썸네일, 상품 이미지, 숏폼 영상을 빠르게 만들어야 한다. 외주 제작은 비용과 수정 리드타임이 크고, A/B 테스트용 소재 수가 부족하다.

구분	내용
핵심 AI 스펙	node-based generation workflow · brand LoRA/style consistency · image-to-video · ControlNet/IP-Adapter · batch generation & review queue
구현 스택	ComfyUI · FLUX · SD3.5 · Wan2.2 · CogVideoX · LTX-Video · ControlNet/IP-Adapter · TTS/lip-sync

CASE STUDY — ANONYMIZED

종합 이커머스 플랫폼 C 사 — 캠페인 콘텐츠 팩토리

산업	규모	도입 범위	GPU 구성
이커머스 (종합 쇼핑 플랫폼)	활성 SKU 약 45 만 개 · 월 캠페인 200+건	배너·상세페이지·SNS 숏폼 대량 생성	이미지 L40S/A100 · 비디오 H100 온디맨드

도입 배경 · 과제

C 사 마케팅팀은 디자이너 12 명과 외주사로 월 3,000 장 수준의 소재를 만들었지만, 45 만 SKU 대비 주요 상품 외에는 소재를 만들 여력이 없었다. 외주 수정 한 번에 3~5 일이 걸려 시즌 캠페인 타이밍을 놓치기 일쑤였고, A/B 테스트는 소재 부족으로 사실상 불가능했다.

GPU-Native 실행 전략

상품 DB 와 캠페인 브리프를 입력하면 ComfyUI workflow 가 브랜드 LoRA 와 ControlNet/IP-Adapter 로 톤이 일관된 배너·상세 이미지 후보를 batch 생성하고, 마케터가 review queue 에서 승인한 이미지는 자동으로 숏폼 영상으로 확장된다. 디자이너는 생산이 아닌 크리에이티브 디렉션과 최종 검수에 집중한다.

PIPELINE 상품 DB → prompt template → ComfyUI(FLUX/SD + ControlNet) → review queue → Wan/CogVideoX 숏폼

도입 성과

지표	도입 전	도입 후
소재 제작 단가	장당 외주 기준	82% 절감
월 소재 생산량	약 3,000 장	36,000 장+ (12 배)
수정 리드타임	3~5 일 (외주)	당일 재생성
광고 성과	A/B 테스트 불가	CTR +19% (다변량 테스트)

Key Point — 이미지 생성은 상시 풀, 비디오 생성은 캠페인 시기에만 H100 온디맨드로 확장 — 생성형 미디어 워크로드의 비용 구조를 외주 변동비에서 통제 가능한 인프라 비용으로 전환.

"gcube.ai 는 마케팅 조직이 생성형 콘텐츠 생산을 외주 중심에서 언제 어디서나 확장 가능한 GPU Native Content Factory 로 전환하도록 지원합니다."

4.7 AI QualityOps / LLMOps

AI 품질·비용·안전성을 측정하는 운영 계층

기업 문제

AI 를 배포했지만 실제로 잘 작동하는지 모르는 경우가 많다. 품질, 환각, 비용, 지연시간, 사용자 만족도, 회귀 오류를 지속 측정하지 못하면 운영 AI 는 빠르게 신뢰를 잃는다.

구분	내용
핵심 AI 스펙	trace-based observability · RAG evaluation · LLM-as-judge · prompt regression testing · red teaming · cost/latency monitoring · feedback-to-dataset loop
구현 스택	Arize Phoenix · Langfuse · Ragas · Promptfoo · DeepEval · OpenTelemetry · Grafana/Prometheus

CASE STUDY — ANONYMIZED

IT 서비스 그룹사 D 사 — 사내 챗봇 QualityOps

산업	규모	도입 범위	GPU 구성
IT 서비스 (그룹 SI·클라우드)	사내 챗봇 MAU 약 9,000 명 · 일 3 만 질의	전 질의 trace · 자동 평가 · 회귀 테스트	judge 모델·batch 평가용 L40S/A100 온디멘드

도입 배경 · 과제

D 사는 사내 규정·기술 문서 챗봇을 배포했지만, 오답에 대한 사용자 불만이 접수돼도 원인이 검색 실패인지 모델 환각인지 파악할 수 없었다. 프롬프트나 모델을 개선하면 다른 질의가 조용히 망가지는 회귀 현상이 반복되며 챗봇 신뢰도가 하락하고 있었다.

GPU-Native 실행 전략

모든 질의응답을 Langfuse 로 trace 하고, 야간 GPU 풀에서 Ragas 기반 faithfulness· context precision 과 LLM-as-judge 평가를 batch 로 수행한다. 실패 케이스는 자동으로 개선 backlog 와 회귀 테스트셋에 추가되어, 이후 모든 변경은 Promptfoo 회귀 테스트를 통과해야 배포된다. 품질·비용·지연시간은 Grafana 대시보드로 운영팀과 경영진에 공유된다.

PIPELINE Langfuse trace → Ragas/LLM-judge 평가 → 실패 마이닝 → Promptfoo 회귀 → Grafana 리포트

도입 성과

지표	도입 전	도입 후
환각성 오답률	약 11%	3%대 (자동 평가 기준)
회귀 오류	배포 후 사용자 신고	배포 전 95% 탐지
토큰·GPU 비용	측정 불가	31% 절감 (라우팅 최적화)
품질 개선 주기	분기 1 회 수동	주 단위 자동 루프

Key Point — 평가·회귀 테스트 워크로드를 야간 유휴 GPU 에 배치해 추가 비용 없이 QualityOps 루프 운영. '배포가 끝'이 아니라 '배포가 시작'인 운영 체계로 전환.

"gcube.ai 는 AI 를 배포하는 데서 멈추지 않고, 품질·비용·안전성을 측정하고 개선하는 AI QualityOps 환경을 제공합니다."

4.8 Trust, Safety & AI Governance Layer

규제 산업을 위한 신뢰 기반 AI 운영

기업 문제

금융, 의료, 공공, 제조, 법률 기업은 개인정보, 접근 권한, 감사 로그, 망분리, 정책 위반 응답을 통제해야 한다. AI 가 업무 시스템과 연결될수록 governance layer 가 필수다.

구분	내용
핵심 AI 스펙	PII detection/masking · prompt injection 방어 · output policy validation · role-based retrieval filtering · audit logging · zero-trust AI access
구현 스택	Microsoft Presidio · NeMo Guardrails · LLM Guard · OPA · Llama Guard · private/on-premise deployment

CASE STUDY — ANONYMIZED

수도권 상급종합병원 H 병원 — 환자정보 보호 AI

산업	규모	도입 범위	GPU 구성
의료 (상급종합병원)	병상 1,200 · 의료진·직원 7,000 여 명	진료·행정 문서 AI 활용 전반의 보안 계층	내부망 전용 GPU (의료 LLM/VLM·RAG)

도입 배경 · 과제

H 병원은 진료기록·검사 결과지를 AI 로 활용하려 했으나 환자정보 보호 규정과 내부망 요구사항 때문에 번번이 무산됐다. 개인정보위 실태점검 대응에도 부서별 문서 접근 이력을 수작업으로 취합해 며칠씩 소요됐고, 결국 민감 문서는 AI 활용 대상에서 제외되어 있었다.

GPU-Native 실행 전략

모든 AI 입력 단계에서 Presidio 가 주민번호·환자명 등 PII 를 실시간 탐지·마스킹하고, OPA 정책과 RAG ACL 필터가 직무 권한에 맞는 문서만 검색되도록 통제한다. 응답은 NeMo Guardrails 로 정책 위반 여부를 검증하며, 모든 질의·응답·접근 이력이 감사 로그로 저장된다. 전체 스택은 내부망의 전용 GPU 환경에서 외부 연결 없이 운영된다.

PIPELINE SSO/RBAC → Presidio 마스킹 → OPA policy → RAG ACL filter → Guardrails 출력 검증 → audit log

도입 성과

지표	도입 전	도입 후

PII 자동 마스킹	수동 비식별화	정확도 99%대 실시간 처리
감사 자료 준비	부서별 취합 5 일	감사 로그 조회 반나절
AI 활용 부서	3 개 (비민감 업무만)	27 개 (진료 지원 포함)
민감정보 유출 사고	잠재 리스크 상존	도입 후 0 건

Key Point — private/on-premise 친화 GPU Runtime 으로 '보안 때문에 AI 를 포기'하던 규제 산업의 도입 장벽을 해소. 보안 계층이 강해질수록 AI 활용 범위가 오히려 확대.

"gcube.ai 는 기업이 보안과 규정을 포기하지 않고도 최신 AI 기능을 내부 통제 환경에서 사용할 수 있도록 지원합니다."

4.9 Synthetic Data Flywheel

실패 데이터를 학습 자산으로 바꾸는 지속 개선 루프

기업 문제

기업은 AI 품질을 높이고 싶지만 고품질 학습 데이터가 부족하다. 실제 고객 데이터는 민감하고, 도메인 전문가는 라벨링 시간이 부족하다. 운영 로그와 실패 케이스를 데이터 자산으로 전환하는 루프가 필요하다.

구분	내용
핵심 AI 스펙	synthetic instruction generation · AI feedback + human review · preference dataset 구축 · DPO/ORPO alignment · eval failure mining
구현 스택	Argilla · distilabel · Phoenix/Langfuse 실패 로그 · Unsloth/Axolotl · Promptfoo regression suite

CASE STUDY — ANONYMIZED

국내 대형 로펌 J 사 — 계약 검토 데이터 플라이휠

산업	규모	도입 범위	GPU 구성
법률 (대형 로펌·기업 법무)	변호사 약 350 명 · 연 계약 검토 수만 건	계약 리스크 검토 AI 의 학습 데이터 파이프라인	synthetic 생성·judge·튜닝 A100/H100

도입 배경 · 과제

J 사는 계약 검토 AI 를 만들고 싶었지만 조항별 리스크 라벨과 우수 수정안 데이터가 절대적으로 부족했다. 변호사들의 검토 의견은 이메일과 메모로 흩어져 있었고, 고연차 변호사에게 라벨링을 요청하는 것은 시간당 비용 구조상 불가능에 가까웠다.

GPU-Native 실행 전략

기존 검토 의견과 계약서를 Presidio 로 익명화한 뒤, distilabel 이 조항별 리스크 질문·답변·수정안 후보를 합성 생성한다. 변호사는 처음부터 작성하는 대신 Argilla 에서 생성 결과를 승인·수정만 하면 되고, 이 판단이 preference dataset 으로 축적되어 DPO 튜닝과 회귀 평가에 재투입된다. 평가에서 발견된 실패 유형은 다음 합성 생성의 시드가 된다.

PIPELINE Presidio 익명화 → distilabel 합성 생성 → Argilla 변호사 검수 → DPO/LoRA 튜닝 → Promptfoo 평가

도입 성과

지표	도입 전	도입 후
학습 데이터 구축 비용	변호사 직접 라벨링 기준	74% 절감
검수 처리 속도	건당 작성 25 분	건당 검수 4 분
계약 초안 검토 시간	기준 1.0	60% 단축
리스크 조항 탐지	누락 빈발	재현율 91% (사내 평가셋)

Key Point — 생성·검수·학습·평가를 하나의 GPU 기반 flywheel 로 운영 — 전문가의 시간을 '라벨 생산'이 아닌 '판단 검수'에만 사용해 희소한 도메인 지식을 데이터 자산화.

"gcube.ai 는 운영 중 발생한 실패와 피드백을 다시 학습 데이터로 전환해 기업 AI 가 시간이 지날수록 똑똑해지는 구조를 만들어 드립니다."

4.10 GPU-Native AI Operating Platform

PoC 를 전사 AI 운영 플랫폼으로 확장

기업 문제

부서별 AI PoC 는 늘지만 운영 환경은 제각각이다. 모델 버전, GPU 사용량, 권한, 평가 결과, 비용 정책이 통합되지 않으면 AI 는 전사 플랫폼이 아니라 실험 모음으로 남는다.

구분	내용
핵심 AI 스펙	Kubernetes GPU orchestration · autoscaling/quota · model registry · multi-model serving · workload-aware scheduling · observability/chargeback
구현 스택	Kubernetes + NVIDIA GPU Operator · KServe · Ray Serve · Triton · MLflow · Dify/Open WebUI

CASE STUDY — ANONYMIZED

국내 10 대 그룹 지주사 G 그룹 — 전사 AI Portal & Runtime

산업	규모	도입 범위	GPU 구성
대기업 그룹 (지주사 IT 총괄)	계열사 14 곳 · 그룹 임직원 4 만여 명	그룹 공통 AI Portal·Gateway·GPU 풀	L40S~H100 혼합 풀 · 워크로드별 스케줄링

도입 배경 · 과제

G 그룹은 계열사마다 별도로 GPU 서버를 도입해 챗봇·RAG PoC 를 진행하면서 동일한 모델·보안 검토·인프라 구축이 14 번 중복되고 있었다. PoC 의 80%는 운영 전환 없이 종료됐고, 각 사가 보유한 GPU 는 평균 사용률 40%를 밑돌았다.

GPU-Native 실행 전략

지주사가 그룹 공통 AI Portal 과 API Gateway, 표준 GPU 풀을 제공하고, 계열사는 RAG·Agent·LLM API 를 셀프서비스로 신청·배포·모니터링한다. 모델은 MLflow registry 로 버전 관리되고, 주간 추론·야간 학습·batch 평가가 워크로드 특성에 따라 하나의 GPU 풀에 스케줄링된다. 사용량은 계열사별로 자동 정산된다.

PIPELINE AI Portal → API Gateway → KServe/vLLM/SGLang → MLflow Registry → QualityOps/Governance → chargeback

도입 성과

지표	도입 전	도입 후
PoC→운영 전환율	약 20%	65%
모델 배포 리드타임	평균 6 주	3 일 (셀프서비스)
그룹 GPU 사용률	평균 38%	71% (통합 스케줄링)
중복 인프라 비용	계열사별 개별 구축	공통 풀 전환으로 연 수십억 원 절감

Key Point — 단일 GPU 대여가 아니라 그룹 전체 AI 운영 플랫폼의 runtime layer 로 포지셔닝 — 계열사가 늘어날수록 단위 비용이 낮아지는 규모의 경제 실현.

"gcube.ai 는 기업이 여러 AI PoC 를 하나의 Enterprise AI Operating Platform 으로 통합할 수 있는 GPU-Native 기반을 제공합니다."

SECTION 05

gcube.ai GPU 라인업별 권장 워크로드

5.1 — GPU Class 정의

GPU Class	예시 GPU	일반적 특성	적합한 방향
Entry / Dev	RTX 4090 급 · L4 급	개발·프로토타입, 경량 추론	PoC · embedding · small model
Inference	RTX 5090 급 · A10 급	비용 효율 추론, 이미지 생성	7B~14B LLM · RAG · ComfyUI
Training / Balanced	A100 80GB 급	학습·추론 균형, 대형 VRAM	LoRA · 32B~70B serving · VLM
High-Performance	H100/H200 급	고성능 학습·추론, FP8 최적화	대규모 LLM · video · high-throughput
Frontier	B200/GB200 급	Blackwell 세대 초고성능	Enterprise AI Platform · frontier

5.2 — 워크로드별 권장 GPU

워크로드	권장 Class	VRAM	권장 스택	설계 포인트
RAG 임베딩	Entry/Inference	16~48GB	BGE-M3 · E5 · TEI	batch 처리·벡터 DB I/O
RAG 리랭킹	Inference/Balanced	24~80GB	bge-reranker	latency·품질 균형
7B~14B LLM 서빙	Inference/Balanced	24~80GB	vLLM · SGLang	continuous batching·양자화
32B~70B LLM 서빙	Balanced/High-Perf	80GB+ multi	vLLM TP · TensorRT-LLM	tensor parallel·KV cache
100B+ 대형 모델	High-Perf/Frontier	multi-GPU	TensorRT-LLM · cluster	network fabric·sharding
LoRA/QLoRA 튜닝	Balanced/High-Perf	48~80GB+	Unsloth · LLaMA Factory	데이터 품질·eval 우선
Full fine-tuning	High-Perf/Frontier	multi-node	Axolotl · DeepSpeed · FSDP	비용·데이터 규모 검토
Document AI/OCR	Inference/Balanced	24~80GB	PaddleOCR · Qwen-VL	batch queue·schema 검증
이미지 생성	Inference/Balanced	24~80GB	ComfyUI · FLUX · SD3.5	workflow·review queue
비디오 생성	High-Perf/Frontier	80GB+	Wan · CogVideoX · LTX	긴 처리시간·batch 스케줄
AI Agent Runtime	Inference/High-Perf	규모별	LangGraph · SGLang	tool latency·trace·승인
QualityOps 평가	Inference/Balanced	24~80GB	Phoenix · Ragas · Promptfoo	야간 batch 평가
Enterprise Platform	Balanced~Frontier	mixed pool	KServe · Ray Serve · MLflow	GPU pool 분리·quota

SECTION 06

Industry Playbooks

산업	핵심 적용 영역	연결 Case Study
금융	Private AI Gateway 통제 · 약관/규정 RAG · PII 마스킹·감사 로그 · 상담 품질 회귀 테스트	4.2 증권사 S 사 · 4.5 보험사 L 사
의료	환자정보 비식별화 · 검사 결과지 Document Intelligence · 가이드라인 RAG · 내부망 배포	4.8 상급종합병원 H 병원
제조	장비 매뉴얼 Knowledge AI · 검사성적서 OCR/VLM · 장애 대응 에이전트 · 멀티모달 QA	4.4 장비 운영사 K 사 · 4.3 제조사 M 사
공공·교육	민원 문서 검색 AI · 보안망 내부 지식 포털 · 교육 콘텐츠 생성 · 정책 준수 응답 관리	4.8 · 4.10 (보안·플랫폼 모델 준용)
이커머스·마케팅	Generative Content Factory · 상품 설명/배너/숏폼 자동 생성 · VOC · 브랜드 톤 튜닝	4.6 이커머스 C 사 · 통신사 T 사
법률·전문서비스	계약 리스크 검토 AI · 판례/계약 Knowledge AI · 요약·수정안 생성 · 전문가 검수 플라이휠	4.9 로펌 J 사

Technical Appendix

10.1 — 구현 스택 후보

영역	Enterprise 용어	구현 후보
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

SECTION 07

Final Message

2026 년 기업 AI 의 본질은 "AI 를 도입했다"가 아니라 "AI 를 운영할 수 있다"에 있다. 기업은 모델을 고르는 단계를 넘어, 모델·지식·에이전트·품질·보안·GPU 인프라를 하나의 실행 체계로 묶어야 한다. Sovereign AI Stack 은 기업이 AI 실행권을 되찾는 전략이다. 그리고 gcube.ai 는 이 전략을 실제 업무 환경에서 구현하기 위한 GPU-Native AI Runtime 이다.

"외부 API 를 호출하는 기업에서, 자체 AI 서비스를 운영하는 기업으로 전환하십시오."

gcube.ai — GPU-Native AI Runtime for Sovereign Enterprise AI

START YOUR SOVEREIGN AI JOURNEY

가성비 RTX 5090 부터, 시장에서 구하기 힘든 B200 까지

gcube.ai 는 워크로드와 예산에 맞는 GPU 를 즉시 제공합니다 — PoC 용 단일 GPU 부터 Frontier 급 전용 클러스터까지.

STEP 1 — GPU 라인업 매칭	STEP 2 — GPU 아키텍처 설계	STEP 3 — 전용 GPU 풀 구축
RTX 5090 급 가성비 GPU 부터 H200·B200 최상위 라인업까지 워크로드에 맞게 제안합니다.	워크로드별 GPU Class·VRAM·네트워크 구성을 설계합니다.	PoC 부터 전사 Runtime 까지 단계적으로 확장합니다.

GPU 공급 및 사업 문의

e-mail : hum@data-alliance.com

Tel : 02-6354-3282

www.gcube.ai

서비스이용문의, 사업제휴문의, 공급 문의 등 모든 문의가 가능합니다.

언제든 문의 주시면, 확인 후 안내드리겠습니다.