

# Sovereign AI Stack 2026

## 기업 AI 주권을 위한 GPU-Native 실행 전략

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

발행 주체: gcube.ai

문서 버전: v0.2 리브랜딩 개정본

작성일: 2026-07-02

대상 독자: CIO, CTO, CDO, AI Transformation 리더, 보안/인프라 책임자, 현업 자동화 리더

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 이다.

## Cover Copy

*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 있다. 이 백서는 최신 AI 스펙을 기업 업무에 적용하기 위한 Sovereign AI Stack 프레임워크와 GPU-Native 실행 아키텍처를 제시한다.

## 1. Executive Brief

생성형 AI 의 첫 번째 물결은 외부 API 와 SaaS 도구를 통해 시작되었다. 많은 기업은 ChatGPT, 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 은 “오픈소스 도구 모음” 이 아니다. 이는 기업이 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 의 역할

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.

2. The Enterprise AI Shift

2.1 2026 년 기업 AI 의 5 대 병목

| 병목      | 기존 증상                   | Sovereign AI Stack 관점의 해결 방향                                   |
|---------|-------------------------|----------------------------------------------------------------|
| 데이터 보안  | 민감 데이터를 외부 API 에 넣기 어려움 | Private AI Gateway, on-premise/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, Enterprise Knowledge AI          |
| 운영 품질   | 배포 후 품질·환각·지연시간 측정 불가   | AI QualityOps, tracing, eval, regression test                  |
| 확장성     | 부서별 PoC 가 플랫폼으로 연결되지 않음 | Enterprise AI Operating Platform, ModelOps, governance         |

2.2 “오픈소스” 가 아니라 “Enterprise-Deployable AI”

백서의 구현 스택은 vLLM, SGLang, LlamaIndex, LangGraph, ComfyUI, Phoenix, Presidio, KServe

등 개방형 생태계를 기반으로 한다. 그러나 전면 메시지는 “오픈소스”가 아니라 다음 언어로 정의한다.

| 일반 표현      | 백서 표현                       |
|------------|-----------------------------|
| 오픈소스 AI    | Open AI Ecosystem           |
| 오픈소스 모델    | Open-Weight Frontier Models |
| 자체 구축      | Private AI Runtime          |
| 사내 ChatGPT | Private AI Gateway          |
| RAG 챗봇     | Enterprise Knowledge AI     |
| AI 에이전트    | Agentic Workflow Engine     |
| 모델 평가      | AI QualityOps               |
| 보안 필터      | Trust & Governance Layer    |
| GPU 서버     | 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                    | 역할                         | 대표 기술                                          |
|--------------------------|----------------------------|------------------------------------------------|
| Application Layer        | 직원·고객·업무 앱이 AI 를 사용하는 접점   | AI Portal, Chat UI, API Client, Workflow UI    |
| Agent Layer              | 목표를 계획하고 도구를 호출하는 업무 실행 계층 | LangGraph, CrewAI, AutoGen, MCP                |
| Knowledge Layer          | 사내 문서와 데이터를 검색 가능한 지식으로 변환 | LlamaIndex, Haystack, Milvus, Qdrant, Weaviate |
| Model Layer              | 기업 업무를 이해하고 생성·분석하는 모델 계층  | Llama, Qwen, Mistral, Gemma, DeepSeek 계열       |
| QualityOps Layer         | 품질, 비용, 지연시간, 회귀 테스트를 운영   | Phoenix, Langfuse, Ragas, Promptfoo            |
| Trust & Governance Layer | 개인정보, 권한, 감사, 안전 정책        | Presidio, NeMo Guardrails, OPA, LLM Guard      |
| GPU-Native Runtime       | 학습·추론·문서·비디오·에이전트를 실행      | gcube.ai GPU Cloud, vLLM, SGLang, KServe       |

#### 3.2 핵심 원칙

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.

## 4. 10 Strategic AI Workloads for 2026

### 4.1 Domain-Adaptive Intelligence

#### 기업 데이터를 이해하는 전용 언어지능

기존 표현: LLM 파인튜닝

전략적 의미: 범용 모델을 기업의 언어, 정책, 도메인 판단 기준에 맞게 적응시키는 계층

#### 기업 문제

범용 LLM은 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.

#### 최신 AI 스펙

- LoRA/QLoRA 기반 parameter-efficient fine-tuning
- DPO/ORPO 기반 선호도 정렬
- instruction dataset 과 preference dataset 분리 운영
- 한국어·다국어 도메인 데이터셋 기반 튜닝
- 평가셋 기반 regression test 후 배포

#### 구현 스택

- Unsloth: 고속 LoRA/QLoRA 학습
- LLaMA Factory: WebUI/CLI 기반 모델 튜닝
- Axolotl: 대규모 실험과 세밀한 학습 설정
- PEFT, TRL, bitsandbytes, FlashAttention
- vLLM/SGLang 기반 튜닝 모델 서빙

## Deep Case Scenario — 고객센터 도메인 상담 LLM

| 항목           | 내용                                                                                     |
|--------------|----------------------------------------------------------------------------------------|
| 산업           | 커머스·금융·통신 고객센터                                                                         |
| 기업 문제        | 상담사는 매일 반복 문의에 대응하지만, 범용 챗봇은 회사 정책·상품명·금지 표현을 정확히 지키지 못함                               |
| 기존 방식        | FAQ 챗봇 + 상담사 수동 응대. 정책 변경 시 롤 업데이트 필요                                                  |
| AI 적용 방식     | 우수 상담 로그, FAQ, 환불/교환 정책, 금지 표현 가이드를 instruction/preference dataset 으로 변환 후 LoRA/DPO 튜닝 |
| 기술 스택        | Presidio 비식별화 → Argilla 검수 → Unsloth/LLaMA Factory 튜닝 → Promptfoo 평가 → vLLM 서빙         |
| GPU 워크로드     | 데이터 정제는 CPU/GPU 혼합, LoRA 학습은 A100/H100 급, 운영 추론은 L40S/A100/H100 급                      |
| 예상 KPI       | 평균 응답 작성 시간 30~60% 단축, 재문의율 감소, 상담 품질 편차 축소, 외부 API 비용 절감                              |
| gcube.ai 포인트 | 민감 상담 데이터 외부 전송 없이 전용 GPU 환경에서 학습·평가·서빙을 통합 운영                                         |

### gcube.ai 메시지

*gcube.ai 는 기업이 자신의 데이터로 전용 언어지능을 만들고, 외부 API 가 아닌 Private AI Runtime 에서 안전하게 운영하도록 지원한다.*

## 4.2 Private AI Gateway

### 외부 API 의존을 대체하는 사내 AI API 계층

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

### 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.

### 최신 AI 스펙

- OpenAI-compatible API
- continuous batching, prefix caching, KV cache optimization
- speculative decoding
- FP8/NVFP4 등 최신 GPU 최적화 양자화
- model routing 과 fallback
- API quota, chargeback, tenant isolation

### 구현 스택

- vLLM, SGLang, TensorRT-LLM, TGI
- LiteLLM 또는 API Gateway
- Prometheus/Grafana, OpenTelemetry
- Phoenix/Langfuse tracing

### Deep Case Scenario — 금융사 Private AI Gateway

|              |                                                                                       |
|--------------|---------------------------------------------------------------------------------------|
| 항목           | 내용                                                                                    |
| 산업           | 금융·보험                                                                                 |
| 기업 문제        | 직원들이 외부 LLM 을 개별적으로 사용하면서 고객 정보 입력 위험과 비용 관리 문제가 발생                                   |
| 기존 방식        | 부서별 SaaS AI 계정, 사용량·데이터 이동 추적 어려움                                                     |
| AI 적용 방식     | 사내 표준 AI Gateway 를 만들고, 모든 LLM 요청을 권한·마스킹·로깅·비용 정책을 거쳐 전용 모델로 라우팅                     |
| 기술 스택        | API Gateway/LiteLLM → Presidio → vLLM/SGLang → Phoenix/Langfuse → Grafana             |
| GPU 워크로드     | 7B~14B 모델은 L40S/A100, 32B~70B 모델은 A100/H100 tensor parallel, 고성능 추론은 H100/H200/B200 급 |
| 예상 KPI       | 외부 API 호출 감소, 부서별 비용 가시화, 민감 데이터 외부 전송 차단, p95 latency 관리                             |
| gcube.ai 포인트 | 전용 GPU 풀과 API 계층을 결합해 기업 내부 표준 AI 인프라로 운영                                             |

### gcube.ai 메시지

외부 AI API 를 통제 불가능하게 호출하는 구조에서 벗어나, gcube.ai 위에 기업 전용 Private AI Gateway 를 구축할 수 있다.

## 4.3 Agentic Workflow Engine

### 판단하고 실행하는 자율 업무 엔진

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

### 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.

### 최신 AI 스펙

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

### 구현 스택

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

### Deep Case Scenario — B2B 영업 제안서 에이전트

| 항목       | 내용                                                                                                  |
|----------|-----------------------------------------------------------------------------------------------------|
| 산업       | B2B SaaS·제조·IT 서비스                                                                                  |
| 기업 문제    | 영업 담당자가 CRM 조회, 고객 리서치, 제안서 초안, 후속 메일 작성에 많은 시간을 사용                                                 |
| 기존 방식    | 사람이 CRM/웹/문서 템플릿을 오가며 수동 작성                                                                         |
| AI 적용 방식 | 에이전트가 CRM 을 조회하고, 산업별 pain point 를 검색하고, 사내 제안서 템플릿과 성공 사례를 RAG 로 불러와 제안서 초안 작성. 발송 전 사람 승인         |
| 기술 스택    | LangGraph planner → CRM API tool → Enterprise Knowledge AI → LLM draft → human approval → audit log |
| GPU 워크로드 | 다수의 짧은 LLM 호출과 RAG 호출. 동시 사용자 증가 시 L40S/A100/H100 기반 autoscaling                                    |
| 예상 KPI   | 제안서 초안 작성 시간 50~80% 단축, 후속 메일                                                                       |

|              |                                                             |
|--------------|-------------------------------------------------------------|
|              | 리드타임 감소, 영업 생산성 증가                                          |
| gcube.ai 포인트 | 에이전트 실행 중 발생하는 추론·검색·문서 생성 워크로드를 전용 GPU Runtime 에서 안정적으로 처리 |

## gcube.ai 메시지

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

## 4.4 Enterprise Knowledge AI

### 사내 문서를 답변 가능한 기업 지식으로 전환

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

### 기업 문제

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

### 최신 AI 스펙

- hybrid search: vector + BM25
- reranker 기반 검색 품질 개선
- metadata/ACL filtering
- citation grounding
- Graph RAG / multi-hop retrieval
- query rewriting and decomposition

### 구현 스택

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

## Deep Case Scenario — 제조 장비 장애 대응 Knowledge AI

| 항목           | 내용                                                                                                    |
|--------------|-------------------------------------------------------------------------------------------------------|
| 산업           | 제조·반도체·장비 운영                                                                                          |
| 기업 문제        | 장비 장애 시 매뉴얼, 과거 정비 이력, 부품 문서를 현장 엔지니어가 직접 찾아야 함                                                       |
| 기존 방식        | 숙련자 의존, 파일 서버 검색, 전화 문의                                                                               |
| AI 적용 방식     | 장비 매뉴얼, 장애 코드, 정비 이력, 부품 카탈로그를 인덱싱하고, 질문에 대해 원문 근거와 조치 순서를 제시                                         |
| 기술 스택        | Docling/PaddleOCR → chunking/metadata → Qdrant/Milvus → BGE-M3 + reranker → LLM answer with citations |
| GPU 워크로드     | 임베딩/리랭킹은 L4/L40S/A100, 답변 생성은 L40S/A100/H100, 대규모 동시 질의는 autoscaling                                  |
| 예상 KPI       | 장애 대응 시간 단축, 숙련자 문의 감소, 정비 절차 표준화, 다운타임 감소                                                            |
| gcube.ai 포인트 | 문서 인덱싱·검색·리랭킹·추론을 하나의 GPU 인프라에서 통합 운영                                                                 |

## gcube.ai 메시지

*gcube.ai 는 기업의 내부 문서를 검색 가능한 지식이 아니라, 바로 업무 답변으로 연결되는 Enterprise Knowledge AI 로 전환한다.*

## 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.

#### 최신 AI 스펙

- layout-aware OCR
- table extraction
- key-value extraction
- PDF-to-Markdown/JSON conversion

- VLM 기반 문서·차트 이해
- human validation workflow

구현 스택

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

Deep Case Scenario — 보험 청구 문서 자동 처리

| 항목           | 내용                                                                                               |
|--------------|--------------------------------------------------------------------------------------------------|
| 산업           | 보험·의료·금융 백오피스                                                                                    |
| 기업 문제        | 보험 청구서, 영수증, 진단서, 검사 결과지를 사람이 수작업으로 입력                                                           |
| 기존 방식        | OCR 일부 + 사람이 검수/입력. 표와 스캔 품질이 낮으면 오류 증가                                                          |
| AI 적용 방식     | 문서 유형 분류 → OCR/VLM 구조 추출 → JSON schema 검증 → 이상치 human review → 업무 시스템 입력                         |
| 기술 스택        | PaddleOCR/Docling/olmOCR → Qwen-VL validation → schema parser → human review UI → ERP/보험 시스템 API |
| GPU 워크로드     | 대량 OCR/VLM batch 처리에는 A100/H100, 경량 OCR은 L40S 급, 피크 처리량은 GPU queue 운영                            |
| 예상 KPI       | 문서 1건 처리 시간 단축, 입력 오류율 감소, 백오피스 처리량 증가, RAG 품질 향상                                                |
| gcube.ai 포인트 | 비정형 문서를 AI-ready 데이터로 바꾸는 전처리 GPU 파이프라인 제공                                                       |

gcube.ai 메시지

*gcube.ai 는 문서AI, VLM, OCR, RAG 인덱싱을 결합해 기업의 비정형 문서를 즉시 활용 가능한 AI 자산으로 전환한다.*

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.

## 최신 AI 스펙

- node-based generation workflow
- brand LoRA / style consistency
- image-to-video, video-to-video
- ControlNet/IP-Adapter 기반 구도·인물·제품 제어
- batch generation and review queue

## 구현 스택

- 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 — 이커머스 캠페인 콘텐츠 팩토리

| 항목           | 내용                                                                                                 |
|--------------|----------------------------------------------------------------------------------------------------|
| 산업           | 이커머스·브랜드 마케팅                                                                                       |
| 기업 문제        | 상품 수가 많아 배너, 상세페이지 이미지, SNS 숏폼을 충분히 제작하지 못함                                                        |
| 기존 방식        | 디자이너와 외주사가 주요 상품 위주로 수동 제작                                                                         |
| AI 적용 방식     | 상품 DB 와 캠페인 브리프를 기반으로 ComfyUI workflow 가 이미지 후보를 생성하고, 승인된 이미지를 숏폼 영상으로 확장                         |
| 기술 스택        | 상품 DB → prompt template → ComfyUI → FLUX/SD + ControlNet/IP-Adapter → Wan/CogVideoX → review queue |
| GPU 워크로드     | 이미지 생성은 L40S/A100, 고해상도·대량 batch 는 A100/H100, 비디오 생성은 H100/H200/B200 급 권장                          |
| 예상 KPI       | 소재 제작 단가 절감, 캠페인 소재 수 증가, 수정 리드타임 단축, CTR/CVR 개선                                                   |
| gcube.ai 포인트 | 생성형 미디어 워크로드를 필요할 때만 GPU 온디맨드로 확장                                                                  |

## gcube.ai 메시지

gcube.ai 는 마케팅 조직이 생성형 콘텐츠 생산을 외주 중심에서 GPU-Native Content Factory 로 전환하도록 지원한다.

## 4.7 AI QualityOps / LLMOps

### AI 품질·비용·안전성을 측정하는 운영 계층

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

### 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.

### 최신 AI 스펙

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

### 구현 스택

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

### Deep Case Scenario — 사내 RAG 챗봇 QualityOps

| 항목    | 내용                                      |
|-------|-----------------------------------------|
| 산업    | 전 산업 공통                                 |
| 기업 문제 | 사내 문서 챗봇이 틀린 답을 해도 운영팀이 원인과 빈도를 파악하지 못함 |
| 기존 방식 | 사용자 불만 접수 후 수동 로그 확인                    |

|              |                                                                                                     |
|--------------|-----------------------------------------------------------------------------------------------------|
| AI 적용 방식     | 모든 질의응답을 trace 하고, RAG faithfulness/context precision, 비용, latency 를 자동 평가. 실패 케이스는 개선 backlog 로 전환 |
| 기술 스택        | Langfuse/Phoenix trace → Ragas eval → Promptfoo regression → Grafana dashboard → Argilla review     |
| GPU 워크로드     | 평가용 judge model, reranking, batch evaluation 에 L40S/A100/H100 사용                                    |
| 예상 KPI       | 환각률 감소, 회귀 오류 사전 탐지, 비용 최적화, 품질 개선 주기 단축                                                            |
| gcube.ai 포인트 | 운영 중인 AI 서비스의 품질 평가와 재학습 루프를 GPU 환경에서 반복 실행                                                         |

## gcube.ai 메시지

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

## 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.

#### 최신 AI 스펙

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

#### 구현 스택

- Microsoft Presidio
- NeMo Guardrails, Guardrails AI, LLM Guard

- OPA(Open Policy Agent)
- Llama Guard 계열
- private deployment / on-premise integration

### Deep Case Scenario — 의료기관 환자정보 보호 AI

|              |                                                                                    |
|--------------|------------------------------------------------------------------------------------|
| 항목           | 내용                                                                                 |
| 산업           | 의료·병원                                                                              |
| 기업 문제        | 환자정보를 포함한 문서를 AI에 활용하고 싶지만<br>개인정보보호와 내부망 요구사항이 강함                                 |
| 기존 방식        | 민감 문서는 AI 도입 대상에서 제외하거나, 사람이<br>수동 비식별화                                            |
| AI 적용 방식     | 입력 단계에서 PII를 탐지·마스킹하고, 권한 있는<br>사용자만 문서 검색 가능. 모든 응답과 접근 이력을<br>감사 로그로 저장          |
| 기술 스택        | SSO/RBAC → Presidio → OPA policy → RAG<br>ACL filter → NeMo Guardrails → audit log |
| GPU 워크로드     | 민감 문서 RAG와 의료 도메인 LLM/VLM 추론을<br>전용 GPU에서 실행                                       |
| 예상 KPI       | 민감정보 노출 위험 감소, 감사 대응 시간 단축,<br>내부망 AI 활용 범위 확대                                     |
| gcube.ai 포인트 | private/on-premise 친화 GPU Runtime으로<br>규제 산업의 AI 도입 장벽 완화                          |

### gcube.ai 메시지

gcube.ai는 기업이 보안과 규정을 포기하지 않고도 최신 AI 기능을 내부 통제 환경에서 사용할 수 있게  
한다.

## 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. 따라서 운영 로그와 실패 케이스를 데이터 자산으로 전환하는 루프가  
필요하다.

## 최신 AI 스펙

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

## 구현 스택

- Argilla, distilabel
- Phoenix/Langfuse/Ragas 실패 로그
- Unsloth/Axolotl/LLaMA Factory
- Promptfoo regression suite

## Deep Case Scenario — 법률 계약 검토 데이터 플라이휠

| 항목           | 내용                                                                                                  |
|--------------|-----------------------------------------------------------------------------------------------------|
| 산업           | 법률·전문서비스·기업 법무팀                                                                                     |
| 기업 문제        | 계약 검토 AI 를 만들고 싶지만 조항별 리스크 라벨과 우수 수정안 데이터가 부족                                                       |
| 기존 방식        | 변호사가 계약서를 수동 검토하고 검토 의견이 문서로 흩어짐                                                                    |
| AI 적용 방식     | 기존 검토 의견을 익명화하고, AI 가 조항별 리스크 질문·답변·수정안 후보를 생성. 변호사가 검수해 preference dataset 으로 축적                   |
| 기술 스택        | Presidio 비식별화 → distilabel synthetic data → Argilla human review → DPO/LoRA tuning → Promptfoo eval |
| GPU 워크로드     | synthetic generation, judge evaluation, fine-tuning 에 A100/H100 급 GPU 사용                            |
| 예상 KPI       | 학습 데이터 구축 비용 절감, 계약 검토 초안 시간 단축, 반복 리스크 탐지율 향상                                                      |
| gcube.ai 포인트 | 데이터 생성·검수·학습·평가를 하나의 GPU 기반 flywheel 로 운영                                                           |

## gcube.ai 메시지

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

## 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.

#### 최신 AI 스펙

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

#### 구현 스택

- 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 — 전사 AI Portal & Runtime

| 항목       | 내용                                                                                               |
|----------|--------------------------------------------------------------------------------------------------|
| 산업       | 대기업·공공·그룹사 IT                                                                                    |
| 기업 문제    | 각 부서가 따로 AI PoC 를 만들고, 모델·데이터·GPU·보안 정책이 중복됨                                                     |
| 기존 방식    | 프로젝트별 VM/GPU 할당, 수동 배포, 비용 추적 어려움                                                                |
| AI 적용 방식 | 공통 AI Portal 과 API Gateway 를 제공하고, 각 부서는 RAG/Agent/LLM API 를 표준 방식으로 신청·배포·모니터링                  |
| 기술 스택    | AI Portal → API Gateway → KServe/Ray Serve/vLLM/SGLang → MLflow Registry → QualityOps/Governance |
| GPU 워크로드 | 주간 추론, 야간 학습, batch evaluation, 미디어                                                              |

|              |                                                      |
|--------------|------------------------------------------------------|
|              | 생성 등 워크로드별 GPU scheduling                            |
| 예상 KPI       | PoC→운영 전환율 증가, 모델 배포 리드타임 단축, GPU 사용률 향상, 부서별 비용 가시화 |
| gcube.ai 포인트 | 단일 GPU 대여가 아니라 전사 AI 운영 플랫폼의 runtime layer 로 포지셔닝    |

### gcube.ai 메시지

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

## 5. gcube.ai GPU 라인업별 권장 워크로드 표

주의: 아래 표는 백서 초안용 권장 아키텍처 매트릭스다. 실제 gcube.ai 제공 GPU 라인업, VRAM, 네트워크, 시간당 가격, 가용 리전은 발행 전 확인 후 확정해야 한다.

### 5.1 GPU Class 정의

| GPU Class               | 예시 GPU            | 일반적 특성                   | 적합한 방향                                        |
|-------------------------|-------------------|--------------------------|-----------------------------------------------|
| Entry / Dev GPU         | RTX 4090 급, L4 급  | 개발·프로토타입, 경량 추론          | PoC, embedding, small model test              |
| Inference GPU           | L40S 급, A10/A30 급 | 비용 효율 추론, 이미지 생성         | 7B~14B LLM, RAG, ComfyUI                      |
| Training / Balanced GPU | A100 80GB 급       | 학습·추론 균형, 대형 VRAM        | LoRA, 32B~70B serving, VLM                    |
| High-Performance GPU    | H100/H200 급       | 고성능 학습·추론, FP8 최적화       | 대규모 LLM, high-throughput serving, video       |
| Frontier GPU            | B200/GB200 급      | Blackwell 세대, 초고성능 추론·학습 | 대규모 enterprise AI platform, frontier workload |

### 5.2 워크로드별 권장 GPU

| 워크로드    | 권장 GPU Class    | VRAM 기준  | 권장 스택                 | 사용 예      | 설계 포인트              |
|---------|-----------------|----------|-----------------------|-----------|---------------------|
| RAG 임베딩 | Entry/Inference | 16~48GB+ | BGE-M3, E5, Jina, TEI | 사내 문서 인덱싱 | batch 처리와 벡터 DB I/O |

|                        |                                |                 |                                    |                         |                                   |
|------------------------|--------------------------------|-----------------|------------------------------------|-------------------------|-----------------------------------|
|                        |                                |                 |                                    |                         | 최적화                               |
| RAG 리랭킹                | Inference/<br>Balanced         | 24~80GB+        | bge-reranker,<br>cross-encoder     | 검색 정확도 개선               | latency와 품질 균형                    |
| 7B~14B LLM 서버          | Inference/<br>Balanced         | 24~80GB+        | vLLM, SGLang                       | 사내 챗봇, 요약               | continuous batching, quantization |
| 32B~70B LLM 서버         | Balanced/<br>High-Performance  | 80GB+ multi-GPU | vLLM TP, SGLang, TensorRT-LLM      | 고품질 QA, 법무/의료           | tensor parallel, KV cache         |
| 100B+ 대형 모델 서버         | High-Performance/<br>Frontier  | multi-GPU       | TensorRT-LLM, vLLM cluster         | 고난도 reasoning           | network fabric, sharding          |
| LoRA/QLoRA 파인튜닝        | Balanced/<br>High-Performance  | 48~80GB+        | Unsloth, LLaMA Factory, Axolotl    | 도메인 LLM                 | 데이터 품질과 eval 우선                   |
| Full fine-tuning       | High-Performance/<br>Frontier  | multi-node      | Axolotl, DeepSpeed, FSDP           | 대형 전용 모델                | 비용·데이터 규모 엄격 검토                   |
| Document AI/OCR        | Inference/<br>Balanced         | 24~80GB+        | PaddleOCR, Docling, Qwen-VL        | 계약서, 청구서                | batch queue, schema validation    |
| 이미지 생성                 | Inference/<br>Balanced         | 24~80GB+        | ComfyUI, FLUX, SD3.5               | 광고 소재                   | workflow template, review queue   |
| 비디오 생성                 | High-Performance/<br>Frontier  | 80GB+           | Wan, CogVideoX, LTX-Video          | 숏폼, 버추얼 휴먼              | 긴 처리시간과 batch scheduling          |
| AI Agent Runtime       | Inference/<br>High-Performance | 규모별             | LangGraph, SGLang, vLLM            | 리서치/영업/코딩 에이전트          | tool latency, trace, approval     |
| AI QualityOps 평가       | Inference/<br>Balanced         | 24~80GB+        | Phoenix, Ragas, Promptfoo          | batch eval, judge model | 운영 로그 기반 야간 평가                    |
| Enterprise AI Platform | Balanced~Frontier              | mixed pool      | KServe, Ray Serve, MLflow, Gateway | 전사 AI 포털                | GPU pool 분리와 quota                |

## 5.3 구매·상담 전환용 간단 매트릭스

| 고객 상황              | 추천 시작점                                | 확장 방향                                  |
|--------------------|---------------------------------------|----------------------------------------|
| “사내 문서 챗봇부터 하고 싶다” | RAG + 7B/14B serving on Inference GPU | reranker, QualityOps, ACL filtering 추가 |
| “외부 API 비용이 너무 크다” | Private AI Gateway + vLLM/SGLang      | model routing, caching, 부서별 quota      |
| “상담봇 품질을 우리 회사     | LoRA fine-tuning on                   | DPO, synthetic data flywheel           |

|                              |                                       |                                                |
|------------------------------|---------------------------------------|------------------------------------------------|
| 스타일로 맞추고 싶다”                 | A100/H100                             |                                                |
| “광고 이미지와 샷폼을 대량 제작하고 싶다”     | ComfyUI on L40S/A100                  | 비디오 생성용<br>H100/H200/B200 pool                 |
| “금융/의료라 보안이 제일 중요하다”         | Private deployment + Governance Layer | Presidio, OPA, Guardrails, audit log           |
| “여러 부서 PoC 를 운영 플랫폼으로 묶고 싶다” | 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]

핵심은 모든 AI 요청이 gateway 를 통과하면서 인증, 정책, 마스킹, 모델 라우팅, 비용 기록, trace 를 남기는 것이다. 이를 통해 기업은 외부 API 의존도를 줄이고 내부 표준 AI API 를 제공할 수 있다.

### 6.2 Enterprise RAG Architecture

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

RAG 는 단순 벡터 검색이 아니라 문서 구조화, 권한 필터링, hybrid retrieval, reranking, citation grounding, evaluation 을 포함해야 한다.

### 6.3 Agentic Workflow Architecture

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

업무 에이전트는 완전 자율보다 통제 가능한 자율성이 중요하다. 고위험 액션은 human approval 을 거치고 모든 도구 호출은 감사 가능해야 한다.

### 6.4 QualityOps & Data Flywheel Architecture

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

운영 로그는 비용 데이터이자 품질 데이터이며, 실패 케이스는 다음 학습 데이터가 된다. 이 루프가 쌓일수록 기업 AI 는 도메인에 맞게 진화한다.

## 6.5 Sovereign AI Stack Architecture

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

Sovereign AI Stack 은 모델, 지식, 에이전트, 품질, 보안, GPU Runtime 을 하나의 계층형 운영 체계로 연결한다.

## 7. Industry Playbooks

### 금융

- Private AI Gateway 로 외부 API 사용 통제
- RAG 기반 상품 약관·규정 QA
- PII 마스킹과 감사 로그
- 상담 품질 평가와 회귀 테스트

### 의료

- 환자 정보 비식별화
- 검사 결과지 Document Intelligence
- 의료 가이드라인 기반 RAG
- 내부망/private deployment

### 제조

- 장비 매뉴얼 Knowledge AI
- 검사성적서 OCR/VLM 처리
- 장애 대응 에이전트
- 현장 작업자용 멀티모달 질의응답

### 공공·교육

- 민원 문서 검색 AI

- 보안망 기반 내부 지식 포털
- 교육 콘텐츠 생성
- 정책 준수형 AI 응답 관리

## 이커머스·마케팅

- Generative Content Factory
- 상품 설명·배너·숏폼 자동 생성
- 고객 VOC 분석
- 브랜드 톤 도메인 튜닝

## 법률·전문서비스

- 계약 리스크 검토 AI
- 판례·계약서 Knowledge AI
- 문서 요약과 수정안 생성
- 전문가 검수 기반 데이터 플라이휠

# 8. ROI Measurement Framework

## 8.1 비용 절감

월간 비용 절감 = 외부 API 비용 절감 + 외주 비용 절감 + 수작업 인건비 절감 - gcube 운영 비용

## 8.2 시간 절감

절감 시간 = 기존 업무 처리 시간 - AI 도입 후 처리 시간  
 금액 환산 = 절감 시간 × 담당자 시간당 비용 × 월 처리 건수

## 8.3 생산성 증가

생산성 증가율 = AI 도입 후 처리량 / 기존 처리량 - 1

## 8.4 리스크 감소

- 민감정보 외부 전송 차단
- 감사 대응 시간 단축
- 규정 위반 응답 차단
- 모델 변경 시 회귀 오류 사전 탐지

## 8.5 신규 매출 기회

- 고객 응대 속도 개선
- 개인화 콘텐츠와 추천
- AI 기반 신규 서비스 출시
- 제품 출시 리드타임 단축

# 9. Adoption Roadmap

| 단계                                        | 목표               | 기간     | 핵심 산출물                                    |
|-------------------------------------------|------------------|--------|-------------------------------------------|
| Phase 1. AI Workload Discovery            | ROI 가 큰 업무 후보 발굴 | 1~2 주  | use case list, data readiness             |
| Phase 2. Private AI PoC                   | 단일 업무 가치 검증      | 2~6 주  | PoC, baseline KPI                         |
| Phase 3. Department Pilot                 | 부서 단위 운영 검증      | 1~3 개월 | limited production, logs, eval            |
| Phase 4. Production AI Runtime            | SLA 와 보안 갖춘 운영   | 3~6 개월 | Gateway, monitoring, governance           |
| Phase 5. Enterprise AI Operating Platform | 전사 AI 플랫폼화       | 6 개월+  | AI Portal, ModelOps, GPU pool, QualityOps |

# 10. 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 |

## 10.2 발행 전 검증 필요 항목

- gcube.ai 실제 GPU 라인업, VRAM, 가격, 네트워크, 스토리지 옵션
- 각 모델과 도구의 최신 라이선스
- 최신 benchmark 수치: vLLM/SGLang/TensorRT-LLM, FLUX/Wan/CogVideoX 등
- 고객 사례 사용 가능 여부
- 금융·의료·공공 보안 표현의 법무 검토

## 11. Final Message

2026 년 기업 AI 의 본질은 “AI 를 도입했다” 가 아니라 AI 를 운영할 수 있다에 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.*