CLIENT CASE STUDY (TEMPLATE)

GenAI Orchestration 기반 Infra / SRE 운영 고도화 컨설팅

Working Diagnosis & 90-Day Prescription — [Client] / [Industry] / [Scale]

Prepared by GyunAI | Version 4.0 | 2026-01-17 | cyberhaven.co.kr

GENERAL FRAMEWORK (PUBLIC)

인프라/플랫폼 운영 효율화 컨설팅

Working Diagnosis & 90-Day Prescription (Anonymized Template)

Prepared by GyunAI | Version 4.0 | 2026-01-17 | cyberhaven.co.kr

Table of Contents

문서 구성(섹션 기반) — 고객/수치/스택은 토큰 치환

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

Key Takeaway: 공개용은 '구조/방법론/산출물' 중심으로, 고객 식별 요소는 전부 토큰화한다.

01 Context & Objectives

고객 의뢰 배경 · 목표 · 범위 정의

Engagement Context

Section 01 | 고객 의뢰 배경(익명)

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

핵심 Pain Point

표준 부재로 운영 복잡도 증가
온콜 피로/MTTR 증가, 재발률 악화
태깅/오너십 부재로 비용 누수 지속
감사/증적 대응이 특정 시점에 폭증

현재 제약(가정)

Cloud-First / Kubernetes 기반
팀 간 도구/프로세스 편차 존재
장애 대응이 '사람' 의존도가 높음
관측성 비용(로그/스토리지) 증가

90일 목표

재현 가능한 운영 표준 확립
SLO 기반 운영(알람 규율/런북)
FinOps 루틴 + 비용 가드레일
증적 자동 패키징으로 리드타임 단축

Key Takeaway: '문제'가 아니라 '운영 가능한 상태'(Standard + SLO + Cost Ownership)로 정의한다.

Executive Summary

Section 01 | 90일 내 '운영 가능한 상태'로 전환

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

Platform 표준화

Landing Zone + IaC 모듈
표준 템플릿(셀프서비스)
변경 리드타임 감소

SRE 운영체계

SLO/에러버짓 + 알람 규율
런북/자동조치로 Toil 감소
MTTR·재발률 개선

FinOps 내재화

태깅/오너십(Allocation)
이상 과금 탐지/가드레일
최적화/커밋 운영 루틴

핵심 메시지

합의는 회의가 아니라 표준과 운영 루틴으로 고정한다. (RACI, SLO, Cost Allocation, Evidence Automation)

Key Takeaway: 표준·자동화·지표로 마찰을 제거하면 DevEx를 유지하면서 안정성/비용/감사를 동시에 만족한다.

02 Diagnosis Framework

가설 → 검증 질문 → 산출물(Discovery 1~2주)

Working Diagnosis

Section 02 | 가설 → 검증 질문 → 산출물

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

Platform

가설: 계정/네트워크/권한/배포 편차
검증: 계정 구조·VPC 패턴·클러스터 수
검증: IaC 커버리지·표준 템플릿 유무
산출: Landing Zone 청사진 + IaC 모듈

SRE / Observability

가설: 알람 노이즈·원인 미상 비중 높음
검증: MTTR/재발률, 알람 유효율
검증: 포스트모템 품질·런북 유무
산출: SLO/에러버짓, 알람 규율, 자동조치 로드맵

FinOps

가설: 태깅/오너십/가드레일 미흡
검증: 태깅 커버리지, Top10 비용 요인
검증: 미사용 리소스, 로그/스토리지 과금
산출: Allocation 표준 + 이상과금 탐지 + 최적화 백로그

Key Takeaway: Discovery는 검증 질문으로 사실화하여 90일 로드맵을 잠금(Lock)한다.

03 Operating Model

RACI · 운영 루틴 · 책임/소유권 고정

Target Operating Model (RACI)

Section 03 | 소유권을 고정해 갈등과 운영 누수를 줄인다

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

RACI Matrix

영역 | Platform/SRE | Dev Teams | Security | Finance/PM

표준 템플릿 / IaC | A/R | C | C | I
배포 파이프라인 / 릴리즈 | A/R | R | C | I
SLO / 온콜 / 런북 | A/R | C | C | I
관측성 표준 / 알람 품질 | A/R | C | I | I
비용 태깅 / 오너십 | A | R | I | A/R
보안 가드레일 | A | C | A/R | I
증적 패키징 | R | I | A | I

Key Takeaway: RACI는 조직도를 바꾸지 않고도 운영 성과를 바꾸는 '최저 비용'의 레버다.

04 Architecture & Standards

Platform · SRE · Observability · FinOps · Security · GenAI Orchestration

Reference Architecture

Section 04 | Cloud-First + Kubernetes + IaC + Standard CI/CD

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

Landing Zone (Account / Network)

멀티 계정(Prod/Stage/Dev) + Guardrails(SCP 등)
표준 VPC/Ingress/Egress 패턴, 감사 로깅 기본값
SSO/IAM 최소권한 + Break-glass 통제

Kubernetes / Deployment

표준 클러스터(네임스페이스/쿼터/네트워크폴리시)
GitOps 또는 표준 배포 템플릿
Policy 기반 배포 가드레일(OPA/Kyverno 등)

Observability + Automation

로그/메트릭/트레이싱 표준 + 대시보드 운영 규율
알람 품질(노이즈 관리) + 온콜 라우팅
런북/자동조치(Auto-remediation) + 롤백 체계

Security + Evidence

Secrets/이미지 스캔/네트워크 정책 기본값화
IaC로 정책/설정 관리 → 증적 자동 생성
감사 시즌 '폭증' 제거(상시 대응)

Key Takeaway: 운영 표준은 '한 장'의 Reference Architecture로 합의하고, IaC/템플릿으로 강제한다.

GenAI Orchestration Reference Model

Section 04 | LLM을 '도구'로 만드는 운영/통제 레이어

Use Case: [Use Case] | Data: [Sources] | Model: [LLM] | Guardrails: [Policy] | Cost: [Budget]

Data Sources
Docs/SaaS/Logs

Ingestion
ETL + PII Redact

Index/RAG
Embeddings + Vector

Orchestrator
Prompt + Tools

Output
Apps/API/Chat

Policy & Guardrails

RBAC/ABAC, Prompt 정책
DLP/PII 마스킹, 데이터 경계
Tool 사용 허용/차단(Allowlist)
감사 로그/레코드 보존

Evaluation & Quality

Golden set + 회귀 테스트
Hallucination/정확도 측정
Safety/Compliance 체크
사용자 피드백 루프

Observability & FinOps

Latency/Errors/Token usage
Prompt/Tool tracing
Cost budget & alerts
캐시/모델 라우팅 최적화

Key Takeaway: GenAI는 '모델'이 아니라 '오케스트레이션'이 제품이다. (정책/평가/관측성/비용이 핵심)

05 Roadmap & Metrics

30/60/90 실행 플랜 + KPI/산출물

30/60/90 Day Plan + KPI

Section 05 | 실행 플랜과 성과 지표(웹 게시용 요약)

Client: [Client] | Industry: [Industry] | Scale: [Scale] | Period: [Period]

0~30일 (Quick Wins)

베이스라인: MTTR/노이즈/비용 Top10
Idle/미사용 종료, 로그 보관 정책
온콜/포스트모템 템플릿 최소 버전

31~60일 (Standardize)

Landing Zone + IaC 모듈 표준화
표준 CI/CD 템플릿 + 롤백 기준
SLO 도입 + 알람 규율 정착

61~90일 (Automate)

런북 자동화/자동조치(Top5)
FinOps 주간 루틴 + 커밋 최적화
증적 자동 패키징 + 가드레일 내재화

Key Takeaway: 90일은 'Quick Win → 표준화 → 자동화/내재화' 3단으로 끊어야 실행된다.

Reuse Notes

홈페이지 게시/포트폴리오 운용 가이드

Public Mode: ON | Client Identifiers: OFF | Tokens: [Client]/[Industry]/[Scale]/[Period]

편집/재사용 규칙

고객명/산업/규모/수치/도구는 [Token]으로 유지 → '찾기/바꾸기'로 일괄 치환
외부 공개 시: 내부 시스템명/계정 구조/보안 통제 상세는 비식별화 또는 삭제
Case Study는 '문제-접근-산출물-성과(KPI)' 4장 구조로 복제
GenAI Orchestration은 '정책/평가/관측성/비용'을 꼭 포함

Contact

GyunAI (Infra / SRE / Platform / DevSecOps / FinOps / GenAI Orchestration)
Website: cyberhaven.co.kr
Deck: Working Diagnosis & 90-Day Prescription (Anonymized)
[Email] / [Phone] / [LinkedIn]

Key Takeaway: 운용 규칙만 지키면 사례를 빠르게 확장할 수 있다.