로컬 실측 변수 1:1 매핑표 & 사내 맞춤 보고서 제작 매뉴얼

벤더 원격 수집 없이 고객사 BYOS 저장소와 로컬 런타임의 실제 계측치를 그대로 출력하는 구조 안내

■ **데이터 정책: Zero Data Custody**
■ 외부 전송량: 0 Byte (사내 완결)
■ 실측 모듈: dock_telemetry
■ 빌더/반출: CDXInternalReportBuilder

1. 왜 벤더가 데이터를 수집하지 않고, 고객사 사내 환경에서 직접 보고서·원본을 생성하는가?

[100% 고객사 BYOS 실측치 직결]

STEP 1. 사내 로컬 실시간 계측 (0 Byte 외부 전송)

- 에이전트 실행 시 로컬 메모리·BYOS 원장 자동 집계
- summary = dock_telemetry.get_summary()
- 호출 0건이면 0건, 페이로드 증가 시 음수(-) 투명 표기
- 고객사 비공개 LLM 계약 단가($) 임의 추측 배제

STEP 2. 기본·커스텀 템플릿 슬롯({변수명}) 자동 매핑

- 기본 3종 보고서(비용정산 / 런타임운영 / 정보보안)
- register_template() 커스텀 양식도 실측치 자동 주입
- org_info (부서·작성자) + metrics (로컬 실측치) 결합
- 비용 환산은 고객사 재무·기획 실제 계약 단가를 적용

STEP 3. 보고서 저장 및 BYOS 원본 1:1 무가공 반출

- reporter.save_report(id, "./out.html") (.md/.csv/.json)
- BYOS 원본 반출: reporter.export_byos_raw_data()
- SQLite3·JSON 4대 저장소 레코드를 엑셀 CSV로 직출
- 사내 그룹웨어 결재·BI 대시보드·감사 증적 즉시 활용

2. [핵심 레퍼런스] 사내 로컬 실측 조회 변수 <-> 보고서 템플릿 슬롯({변수명}) 1:1 매핑표

[단가 추측 배제 · 순수 로컬 실측치]

구분 (대상 양식)	실측 집계 항목 설명	사내 로컬 데이터 조회 경로	템플릿 슬롯 키 ({변수명})
공통 헤더 (org_info)	보고 부서명 / 작성자 성명·직급	고객사 직접 입력 (org_info 딕셔너리)	{company_name}, {author_name}
공통 헤더 (org_info)	집계 기간 / 운용 환경 / 플랜명	고객사 직접 입력 (예: '2026-10 / 사내 VPC')	{reporting_period}, {active_plan}
① 비용·토큰 정산 (호출·원본토큰)	누적 에이전트 실행 호출 수 (폴백 포함)	summary['total_agent_calls']	{monthly_agent_calls}
① 비용·토큰 정산 (전후 토큰비교)	처리 전 원본 토큰 vs 처리 후 전송 토큰	summary['input_tokens']['raw'] / ['slimmed']	{raw_tokens_total} {slimmed_tokens_total}
① 비용·토큰 정산 (순증감·비율)	부호 포함 순 증감 토큰(+/-) 및 비율(%)	summary['net_tokens_delta'] summary['overall_token_savings_percent']	{saved_tokens_total}
① 비용·토큰 정산 (모듈별 절감량)	모듈별 실측 절감 토큰량 (Arbiter / Slimmer 등)	summary['module_token_savings']	{stage1_arbiter_saved_tokens} {stage2_slimmer_saved_tokens}
① 비용·토큰 정산 (비용 환산 산식)	고객사 자체 LLM 계약 단가 곱셈 산식	고객사 재무·기획팀 실제 계약 단가표 적용	{customer_cost_formula}
② 런타임 운영 (처리율·지연)	실측 정상 처리율(%) 및 평균 지연시간(ms)	summary['completion_rate_pct'] summary['average_latency_ms']	{total_pipeline_runs} {avg_dock_overhead_ms}
② 런타임 운영 (CoT 루프 제어)	<think> 동어반복 루프 조기 절단 건수	summary['cot_guard']['loops_broken']	{cot_guard_loops_detected} {cot_timeouts_prevented}
② 런타임 운영 (기억팩·폴백)	기억팩 실측 회상 지연 / 예외 폴백 건수	summary['average_tokenpack_latency_ms'] summary['fallback_calls']	{tokenpack_recall_latency} {fallback_activations}
⑤ 정보보안 점검 (BYOS 볼트)	사내 로컬 BYOS 볼트 경로 및 핀 카테고리	reporter.cloak.vault_db_file reporter.cloak.explicit_secrets	{byos_storage_location} {cloak_pinned_categories}
⑤ 정보보안 점검 (복원·정적점검)	실측 가명화·복원 건수 / 실시간 AST 스캔	summary['cloak_byos'] verify_open_adapter_static_imports()	{total_masked_outbound_calls} {static_code_audit_result}

3. 맞춤 보고서 작성 및 BYOS 원본 데이터(.csv / .json) 1:1 반출 예제 코드

[사내 서버 즉시 실행 가능]

```
from cdx_dock import dock_telemetry, CDXInternalReportBuilder

# 1) 사내 로컬에서 실측된 텔레메트리 조회 (호출 0건 시 0 출력, 추정 단가 배제)
reporter = CDXInternalReportBuilder(storage_path='./enterprise_vault.db', persist_telemetry=True)

# 2) 사내 맞춤 양식 등록 — {변수명} 슬롯에 로컬 BYOS 실측치가 자동 바인딩됨 (.md / .html / .csv / .json 지원)
reporter.register_template(
    template_id='weekly_team_report',
    title='[주간 운영 요약] AI 에이전트 실측 토큰 및 보안 점검표',
    template_markdown='부서: {company_name} | 호출: {monthly_agent_calls} | 순증감: {saved_tokens_total} | 복원: {local_unmask_restorations}'
)
reporter.save_report('weekly_team_report', './weekly_report.html', use_live_telemetry=True)
reporter.export_byos_raw_data('./byos_raw_passthrough.csv', include_secret_values=False)  # BYOS 원본 1:1 반출
```

[양식 레이아웃 예시 샘플 #1] template_id = 'cfo_roi_settlement_report' (실행 시 당사 실측치 출력)

담 당	팀 장	부서장	임 원
김민수 (승인)	이준호 (승인)	박정훈 (검토)	결재대기

[내부 결재 품의서] AI 에이전트 토큰 절감량(Tokens) 및 절감률(%) 정산서

문서번호: FIN-2026-1031-01 | 집계원칙: 로컬 실측 토큰 증감량만 기재 (금액 환산은 당사 실제 LLM 계약단가 적용)

Ⅰ. 품의 기본 정보 및 비용 환산 기준 (고객사 입력 변수: org_info)

[양식 헤더 예시]

항목명	출력 내용 (레이아웃 예시 데이터)	바인딩 슬롯 키 (org_info)
보고 부서 / 법인명	Global FinTech Corp. (AI 플랫폼실)	{company_name}
작성자 (담당 엔지니어)	김민수 수석엔지니어 (AI 인프라팀)	{author_name}
성과 집계 대상 기간	2026년 10월 01일 ~ 2026년 10월 31일	{reporting_period}
운용 아키텍처 환경	사내 프라이빗 VPC 온프레미스 (외부 벤더 전송 0 Byte)	{deployment_mode}
재무 비용 환산 기준	당사 재무·기획팀 실제 LLM 공급 계약 단가표(1M 토큰당 단가) 별도 적용	{customer_llm_contract_note}

Ⅱ. 월간 토큰 발생량 및 순증감 요약 (레이아웃 예시 — 실사용 시 로컬 실측치 자동 주입)

[단가 추측 배제]

① 처리 전 원본 토큰 발생량

3.12B Tokens [예시]

예시: 월간 1,500,000회 호출 기준

템플릿 변수: {raw_tokens_total}

② CDX 처리 후 실제 전송 토큰

1.48B Tokens [예시]

상용/사내 LLM으로 실제 송출된 토큰

템플릿 변수: {slimmed_tokens_total}

③ 순 절감 토큰량 및 비율(%)

1.64B 절감 (-52.6%)

증가 시 음수(-)/증가율(+) 투명 표기

템플릿 변수: {saved_tokens_total}

Ⅲ. 모듈별 토큰 절감량(Tokens) 및 감축 비율(%) 상세 내역

[dock_telemetry 로컬 실측 매핑]

최적화 단계 및 모듈명	세부 기술 작동 원리	양식 예시 표기 (실행 시 실측치)	템플릿 슬롯 키 (metrics)
1. 도구 스키마 선별 라우팅 (CDXArbiter)	전체 도구 스키마 일괄 주입 대신 질의별 Top-K 관련 도구만 선별 주입하여 입력 스키마 토큰 감축	810,000,000 Tokens (스키마 -93.3% 예시)	{stage1_arbiter_saved_tokens}
2. 페이로드 가역 압축 코덱 (DockSlimmer Mode B)	에이전트 간 전달 JSON의 중괄호·따옴표 구문 중복을 줄인 Mode B 변환으로 입출력 페이로드 감축	460,000,000 Tokens (페이로드 -48.2% 예시)	{stage2_slimmer_saved_tokens}
3. 추론 루프 조기 절단 (CDX-CoT-Guard)	추론형 모델(<think>)의 동어반복 루프를 실시간 감지하여 불필요한 추론 스트림 토큰 조기 절단	210,000,000 Tokens (루프 3,842건 예시)	{stage3_cot_guard_saved_tokens}
4. 장기 기억 중복 제거 (CDX-TokenPack-10K)	전체 대화 이력 반복 재주입 대신 로컬 기억팩에서 핵심 문맥만 선별 회상하여 입력 토큰 감축	160,000,000 Tokens (이력 재주입 감축 예시)	{stage4_5_memory_saved_tokens}
합계 및 자체 비용 산식	고객사별 사용 LLM 모델 및 기업 계약 단가가 상이하므로 우측 산식에 당사 실제 단가를 곱하여 정산함	순절감 토큰량 × [당사 LLM 계약 단가]	{saved_tokens_total} {customer_cost_formula}

■ 집계 안내 (벤더 추정 단가를 배제하고 고객사 로컬 실측 토큰량만 표시하는 이유):

- 본 PDF는 문서 레이아웃 확인을 위한 샘플이며, 실제 SDK에서 use_live_telemetry=True(기본값)로 실행 시 호출 0건이면 '0회 / 0 Tokens'가 정직하게 출력됩니다.
- 고객사의 LLM 모델 구성 및 기업별 할인 단가는 고객사 고유 정보이므로, 사내 실측 토큰 증감량에 당사 실제 계약 단가를 곱하여 정산하실 수 있습니다.

[실행 명령어] reporter.save_report('cfo_roi_settlement_report', './token_settlement.html', use_live_telemetry=True)

[양식 레이아웃 예시 샘플 #2] template_id = 'cto_agent_runtime_report' (실행 시 당사 실측치 출력)

[기술 운영 보고서] 멀티 에이전트 런타임 가동 및 추론 루프 제어 리포트

문서번호: OPS-2026-1031-02 | 점검 대상: LangChain <-> CrewAI <-> Hermes 프로덕션 파이프라인

담 당	팀 장	부서장	임 원
김민수 (승인)	이준호 (승인)	박정훈 (검토)	결재대기

Ⅰ. 엔지니어링 운영 환경 기본 정보 (고객사 입력 변수: org_info)

[기술·플랫폼 조직 명세]

항목명	운영 환경 명세 (예시)	템플릿 변수 키 (org_info)
작성 조직 / 담당자	Global FinTech Corp. (AI 플랫폼실) / 이준호 팀장 (에이전트 아키텍처팀)	{company_name}, {author_name}
운영 점검 기간	2026년 10월 01일 ~ 2026년 10월 31일	{reporting_period}
연동 에이전트 파이프라인	LangChain (SQL 분석) <-> CrewAI (리포트 팀) <-> Nous Hermes (SKILL.md 연동)	{connected_frameworks}
미들웨어 배포 형태	In-Process Python SDK (cdx-dock v2.0) + Kubernetes Docker Sidecar v2.0	{deployment_mode}

Ⅱ. 멀티 에이전트 런타임 가동 및 실측 지연시간 KPI (레이아웃 예시)

[실측 perf_counter 기반]

① 파이프라인 실행 및 정상 처리율

412,800건 (99.98%)

예외 폴백 발생 건수도 가감 없이 집계

템플릿 변수: {total_pipeline_runs}

② 미들웨어 실측 평균 지연시간

0.41 ms [예시]

로컬 인메모리 실측 평균 지연시간 표기

템플릿 변수: {avg_dock_overhead_ms}

③ <think> 동어반복 루프 조기 절단

3,842건 [예시]

장시간 스트림 지연 및 타임아웃 완화

템플릿 변수: {cot_guard_loops_detected}

Ⅲ. 추론 루프 제어(CoT-Guard) 및 장기 기억(TokenPack-10K) 세부 지표

[dock_telemetry.get_summary() 연동]

세부 점검 항목	양식 예시 표기 및 계측 기준	로컬 텔레메트리 경로	템플릿 슬롯 키 (metrics)
파이프라인 실행 및 처리율	총 실행 건수 및 실측 정상 완료 비율(%) 표시	summary['total_agent_calls'] summary['completion_rate_pct']	{total_pipeline_runs}
미들웨어 실측 평균 지연	어댑터·라우터·압축 구간의 실제 측정 평균 지연시간(ms)	summary['average_latency_ms']	{avg_dock_overhead_ms}
<think> 동어반복 루프 감지	추론 스트림 내 반복 고착 패턴 실시간 감지 건수	summary['cot_guard'] ['loops_broken']	{cot_guard_loops_detected}
추론 루프 조기 절단 건수	반복 루프 조기 종료 유도로 장시간 응답 지연을 예방한 건수	summary['cot_guard'] ['loops_broken']	{cot_timeouts_prevented}
절감된 추론 스트림 토큰	루프 조기 절단으로 추가 발생을 막은 추론 토큰 수	summary['cot_guard'] ['tokens_saved']	{cot_wasted_tokens_saved}
TokenPack-10K 회상 속도	로컬 SQLite 기억팩 조회 시 실제 측정된 평균 회상 시간 및 횟수	summary['tokenpack_recalls'] & average_tokenpack_latency_ms	{tokenpack_recall_latency}
Safe Pass-Through 폴백	상위 프레임워크 예외 시 안전 원문 우회가 동작한 실제 건수	summary['fallback_calls']	{fallback_activations}

■ 기술 운영 점검 요약 (예시 문안):

"프론티어 추론 모델(GPT-6 Astra, Claude Mythos 5.2, Gemini 4 Carbon 등) 운용 시 간헐적으로 발생하던 <think> 반복 루프를 조기 제어하고,
별도 외부 벡터 DB 없이 로컬 파일 기반 TokenPack-10K로 다중 프레임워크 간 핵심 문맥을 일관되게 공유 중임을 로컬 실측 지표로 확인함."

[실행 명령어] reporter.save_report('cto_agent_runtime_report', './runtime_ops.csv', use_live_telemetry=True)

[양식 레이아웃 예시 샘플 #3] template_id = 'ciso_security_audit_report' (실행 시 당사 실측치 출력)

담 당	팀 장	부서장	임 원
김민수 (승인)	이준호 (승인)	박정훈 (검토)	결재대기

[정보보안 점검 보고서] 사내 BYOS 기밀 가명화 및 외부 무전송 점검서

문서번호: SEC-2026-1031-03 | 점검 기준: 개인정보보호법 안전조치 · ISMS-P · 사내 보안성 검토 기준

Ⅰ. 정보보안 점검 기본 정보 및 사내 BYOS 볼트 저장 경로 (org_info)

[온프레미스 격리 확인]

점검 항목	보안 통제 내역 (예시)	템플릿 변수 키 (org_info)
점검 부서 / 담당자	Global FinTech Corp. (정보보호팀) / 박정훈 수석 (보안·컴플라이언스 파트)	{company_name}, {author_name}
점검 대상 기간	2026년 10월 01일 ~ 2026년 10월 31일	{reporting_period}
사내 BYOS 볼트 파일 경로	/var/lib/enterprise/cdx_byos/enterprise_vault.db (사내 로컬 SQLite3 파일)	{byos_storage_location}
외부 벤더 원격 전송량	0 Byte (외부 원격 수집 모듈 미포함 · 사내 로컬 직접 집계)	{vendor_data_transmission}

Ⅱ. BYOS 가명화·복원 및 정적 코드 점검 핵심 지표 (레이아웃 예시)

[정보보안·컴플라이언스 검토 지표]

① 외부 벤더 원격 데이터 전송량

0 Byte (무전송)

프롬프트·매핑원장 사내 로컬 보관

템플릿 변수: {vendor_data_transmission}

② 가명화 송출 및 로컬 복원 실측

184,920건 [예시]

실제 치환된 식별자 개수까지 정밀 표기

템플릿 변수: {total_masked_outbound_calls}

③ 오픈 어댑터 실시간 AST 점검

PASS (실측 스캔)

설치된 .py 파일 내 외부 통신 import 검사

템플릿 변수: {static_code_audit_result}

Ⅲ. CDXCloak 기밀 핀 등록 현황 및 로컬 AST 정적 점검 세부 내역

[실시간 로컬 검증]

보안 점검 항목	사내 로컬 실측 출력 내용 (예시)	데이터 출처 (BYOS / AST)	템플릿 슬롯 키 (metrics)
BYOS 등록 기밀 카테고리	VIP_CUSTOMER, SETTLEMENT_ACC, BIZ_REG_NO 등 사내 볼트에 실제 등록된 핀 라벨 및 건수 출력	reporter.cloak.explicit_secrets	{cloak_pinned_categories}
외부 LLM 송출 전 가명화	외부 API 송출 전 [[VAULT_PIN_*]] 토큰으로 실제 치환된 호출 건수 및 치환 식별자 총 개수	summary['cloak_byos'] ['masked_calls']	{total_masked_outbound_calls}
응답 수신 후 로컬 복원	응답 내 가명 토큰이 실제로 존재하여 원본으로 복원된 호출 건수 및 복원 식별자 총 개수 (무토큰 시 미집계)	summary['cloak_byos'] ['unmasked_calls']	{local_unmask_restorations}
마스킹 동작 정책 모드	strict_pin_mode=True 적용 여부 (명시 등록된 기밀 핀만 치환하여 일반 날짜·금액·코드 수치 변조 방지)	reporter.cloak.strict_pin_mode	{false_positive_mutations}
오픈 어댑터(.py) 정적 점검	설치된 오픈 어댑터 .py 파일들을 Python ast 모듈로 실시간 파싱하여 requests/httpx 등 import 유무 점검	verify_open_adapter_ static_imports()	{static_code_audit_result}

■ 데이터 보관 확인 및 BYOS 원본 감사 증적 반출 안내:

- 모든 기밀 매핑 원장(enterprise_vault.db)과 텔레메트리 카운터는 고객사가 지정한 사내 로컬 스토리지(BYOS)에만 격리 보관됩니다.
- 정보보안팀 감사 증적이 필요할 경우 reporter.export_byos_raw_data('./audit_raw.csv')로 BYOS 원본 레코드를 그대로 추출하실 수 있습니다.

[실행 명령어] reporter.save_report('ciso_security_audit_report', './security_audit.html', use_live_telemetry=True)