[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)