[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)