[LAYOUT PREVIEW SAMPLE #1] template_id = 'cfo_roi_settlement_report' (lang='en')

[Token & Cost Settlement] Measured Token Savings & Ratio (%) Report

Doc Ref: FIN-2026-1031-01 | Accounting Basis: Strictly measured token deltas (apply internal LLM rate per 1M tokens)

Prepared	Reviewed	Director	Approved
A. Kim Signed	J. Lee Signed	M. Park Reviewed	Pending

I. Organization Metadata & Cost Conversion Basis (Customer Input: org_info)

[Sign-off Header Preview]

Field	Sample Preview Value	Bound Template Slot (org_info)
Organization / Division	Global FinTech Corp. (AI Platform Division)	{company_name}
Prepared By	Alex Kim, Principal AI Infrastructure Engineer	{author_name}
Reporting Period	2026-10-01 to 2026-10-31 (31 Days)	{reporting_period}
Deployment Environment	On-Premise Private VPC (0 Byte External Vendor Transmission)	{deployment_mode}
FinOps Conversion Note	Apply Finance / FinOps confidential contracted LLM rate per 1M tokens	{customer_llm_contract_note}

II. Measured Monthly Token Volume & Net Delta Summary (Layout Preview)

[Zero Vendor Price Guesswork]

1. Pre-Optimization Raw Tokens

3.12B Tokens [Sample]

Preview: 1,500,000 agent calls

Slot Key: {raw_tokens_total}

2. Post-CDX Outbound Tokens

1.48B Tokens [Sample]

Actual tokens sent to LLM APIs

Slot Key: {slimmed_tokens_total}

3. Measured Net Token Delta (%)

1.64B Saved (-52.6%)

Reports negative delta if expanded

Slot Key: {saved_tokens_total}

III. Per-Module Measured Token Savings & Ratio Breakdown

[dock_telemetry.get_summary()]

Optimization Stage & Module	Technical Mechanism	Sample Preview Value	Template Slot Key (metrics)
1. Selective Schema Routing (CDXArbiter)	Injects only Top-K relevant tool schemas per query instead of broadcasting full tool catalogs	810,000,000 Tokens (Schema -93.3% preview)	{stage1_arbiter_saved_tokens}
2. Lossless Payload Codec (DockSlimmer Mode B)	Compacts repetitive JSON braces/quotes across inter-agent payloads via reversible Mode B	460,000,000 Tokens (Payload -48.2% preview)	{stage2_slimmer_saved_tokens}
3. Reasoning Loop Cutoff (CDX-CoT-Guard)	Detects repetitive <think> reasoning loops in streaming outputs and halts wasted generation	210,000,000 Tokens (3,842 loops preview)	{stage3_cot_guard_saved_tokens}
4. Shared Context Pack (CDX-TokenPack-10K)	Recalls compact specs from local SQLite pack instead of re-sending full chat histories	160,000,000 Tokens (History dedup preview)	{stage4_5_memory_saved_tokens}
Total & FinOps Formula	Multiply measured net saved tokens by your internal contracted LLM rate per 1M tokens	Net Saved Tokens x [Contracted LLM Rate]	{saved_tokens_total} {customer_cost_formula}

Accounting & Transparency Note (Why Zero Vendor Dollar Estimates Are Injected):

- This PDF illustrates document formatting only. In live execution (use_live_telemetry=True), 0 calls output '0 calls / 0 Tokens' with zero sample inflation.
- Because enterprise LLM model mixes and negotiated discount rates are confidential, Finance teams apply their actual rate to measured token deltas.

[Command] reporter.save_report('cfo_roi_settlement_report', './token_settlement.html', use_live_telemetry=True, lang='en')