

pronix.ai

An AI-first business of Pronix Inc.

PLAYBOOK · PNX-PBK-739

The FinOps playbook for LLM + CCaaS spend

How to control LLM and CCaaS spend before it controls you. Token analytics, model routing, license rightsizing, telephony minute optimization and a board-level cost dashboard you can copy directly.

PREPARED FOR

CFO, CIO, Head of FinOps, VP AI Platform

PREPARED BY

pronix.ai FinOps Practice, AI cost & governance

EDITION

Updated Q1 2026 · 15 min read · 28-page PDF

STRATEGY · IMPLEMENTATION · TALENT pronix.ai/resources

EXECUTIVE SUMMARY

The FinOps playbook for LLM + CCaaS spend

How to control LLM and CCaaS spend before it controls you. Token analytics, model routing, license rightsizing, telephony minute optimization and a board-level cost dashboard you can copy directly.

20-40%

Typical LLM spend reduction
from model routing

15-25%

CCaaS license savings from
rightsizing

6

Metrics on the board
dashboard

What you'll take away

- How to instrument token, minute and seat spend across providers
- Model routing patterns — cheap-model-first, cascade, and quality gates
- License rightsizing across CCaaS, CRM, and AI providers
- Telephony minute optimization on Amazon Connect, Genesys and NICE
- The board cost dashboard: 6 metrics that matter
- Guardrails against runaway spend before it hits your invoice

Audience: CFO, CIO, Head of FinOps, VP AI Platform

CONTENTS

What's inside

#	Section	Focus
01	The cost anatomy of enterprise AI	Where the money actually goes across LLM tokens, CCaaS seats, telephony minutes, storage, and observability.
02	Instrumentation	Tagging, attribution, cost-per-conversation and cost-per-outcome — the 4 metrics FinOps and product both trust.
03	Model routing	Cheap-first cascades, quality gates, provider failover, batch vs realtime — with per-provider price tables refreshed qua
04	License rightsizing	Named-user vs concurrent, seasonal peaks, marketplace commitments, and how to hold vendors to their published discounts.
05	Telephony optimization	Regional dial plans, carrier arbitrage, containment x minute-price interaction — usually the largest single line item.
06	The board dashboard	6 metrics: cost per resolution, containment rate, AI margin, model mix, seat utilization, forecast variance.

SECTION 01

The cost anatomy of enterprise AI

Where the money actually goes across LLM tokens, CCaaS seats, telephony minutes, storage, and observability.

In our engagements, the practical work in this section usually breaks down into three deliverables: the board cost dashboard: 6 metrics that matter.

The finance conversation is won or lost on unit economics. Instrument cost-per-resolution, cost-per-decision or cost-per-contained-contact from day one, publish the run-rate weekly, and make trade-offs — model tier, context length, cache strategy — visible to product owners rather than buried in a platform bill.

SECTION 02

Instrumentation

Tagging, attribution, cost-per-conversation and cost-per-outcome — the 4 metrics FinOps and product both trust.

In our engagements, the practical work in this section usually breaks down into three deliverables: how to instrument token, minute and seat spend across providers; model routing patterns — cheap-model-first, cascade, and quality gates.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

WHAT GOOD LOOKS LIKE

Model routing patterns — cheap-model-first, cascade, and quality gates

SECTION 03

Model routing

Cheap-first cascades, quality gates, provider failover, batch vs realtime — with per-provider price tables refreshed quarterly.

In our engagements, the practical work in this section usually breaks down into three deliverables: model routing patterns — cheap-model-first, cascade, and quality gates.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

SECTION 04

License rightsizing

Named-user vs concurrent, seasonal peaks, marketplace commitments, and how to hold vendors to their published discounts.

In our engagements, the practical work in this section usually breaks down into three deliverables: license rightsizing across ccaas, crm, and ai providers.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

WHAT GOOD LOOKS LIKE

Telephony minute optimization on Amazon Connect, Genesys and NICE

SECTION 05

Telephony optimization

Regional dial plans, carrier arbitrage, containment × minute-price interaction — usually the largest single line item.

In our engagements, the practical work in this section usually breaks down into three deliverables: telephony minute optimization on amazon connect, genesys and nice.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

SECTION 06

The board dashboard

6 metrics: cost per resolution, containment rate, AI margin, model mix, seat utilization, forecast variance.

In our engagements, the practical work in this section usually breaks down into three deliverables: the board cost dashboard: 6 metrics that matter.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

WHAT GOOD LOOKS LIKE

Guardrails against runaway spend before it hits your invoice

FREQUENTLY ASKED

Questions from enterprise buyers

Does this work across multiple AI providers?

Yes — the routing patterns assume you use at least two of OpenAI, Anthropic, Bedrock, Azure AI Foundry, Gemini and Kore.ai Agent Platform, and the license section covers the major CCaaS and CRM providers.

Is the board dashboard included as a template?

Yes — the download includes a Looker/Metabase-ready query pack and a Google Slides board template.

TURN THIS INTO A DECISION

Book a 30-minute working session.

Bring your platform, data, seats and delivery constraints. A pronix.ai strategy or CX engineering lead will walk this document through with your team and adapt it — no slideware, no generic playbook.

ENTERPRISE SALES	hello@pronix.ai	pronix.ai/contact
PARTNERSHIPS & BPO	partners@pronix.ai	pronix.ai/company/partners
TALENT SOLUTIONS	talent@pronix.ai	pronix.ai/talent

pronix.ai is an AI-first business of Pronix Inc. All third-party product names — Amazon Connect, Genesys Cloud, NICE CXone, Five9, Kore.ai, Salesforce Agentforce, Microsoft, Google Cloud, AWS, IBM watsonx — are trademarks of their respective owners. Referenced only to describe integration scope; no endorsement is implied.