THOUGHT LEADERSHIP

WINNING THE NEXT COST BATTLE

Enterprise AI spend is accelerating faster than most organizations can govern it. Chain IQ's IT Sourcing team on why the winning move is not negotiating harder on price, but building the right portfolio of AI suppliers.



INDUSTRIES

Cross-Industry | Technology & IT

EXPERTISE

AI Procurement | IT Sourcing | Cost Optimization

AUTHORS

Ivo Claus
Global Head of IT Sourcing, Chain IQ

Joe Sweet
Global Head of Lead Buying, Chain IQ

AT A GLANCE

From software licence to consumption cost, and CIOs are feeling it first

Across industries, AI adoption has become a board-level priority. Unlike software licenses, AI costs are driven by consumption. Every prompt, workflow, agent, and model interaction consumes tokens, compute, storage, and supporting infrastructure. As adoption scales, organizations frequently lack visibility into which workloads generate cost and whether those costs are delivering proportional business value.

Traditional procurement disciplines remain critical, but AI introduces new commercial variables including token pricing, consumption commitments, model flexibility, price protection, workload routing rights, data residency requirements, and future access to new model releases.

In this conversation, Ivo Claus, Chain IQ's Global Head of IT Sourcing, and Joe Sweet, Global Head of Lead Buying, explain why the greatest savings opportunity in AI often comes not from negotiating a lower price, but from avoiding long-term commitments in a market where technology performance improves and pricing continues to fall.

OBJECTIVE

The bill arrives after the decision

The biggest misconception, Claus says, is treating AI as just another software category. AI is fundamentally a consumption business. Every prompt, every chatbot exchange, every autonomous agent and automated workflow draws on tokens, compute, and storage.

As usage scales across thousands of employees, spending can increase faster than budgets, business cases, and governance controls.

The challenge is not simply cost visibility. The challenge is designing a commercial structure that protects the organization if AI prices decline significantly over the next 12-36 months while at the same time usage might be growing exponentially.

CONCRETE USE CASE

AI agreement negotiation for a Chain IQ client



A Chain IQ client was preparing a major enterprise agreement with Anthropic for Claude Code. The supplier proposal was based on a traditional subscription commitment model despite significant uncertainty around future adoption and production demand.

BEFORE

METRIC	VALUE
Supplier	Anthropic (Claude Code)
Contract Value	USD 620,000
Users	500
Commercial Model	Fixed Subscription Commitment
Commitment Profile	Upfront Capacity Commitment

The client faced the risk of committing to future demand before actual enterprise adoption patterns had been validated.

CHAIN IQ APPROACH

Chain IQ:

- Separated enterprise access from consumption economics
- Benchmarked the commercial model against broader market practice
- Challenged supplier assumptions regarding future usage
- Replaced fixed commitments with token-based consumption pricing
- Negotiated phased adoption and governance controls
- Structured the agreement around flexibility rather than volume commitments

AFTER

IMPROVEMENT DELIVERED	OUTCOME
Enterprise Seats	500 enterprise seats secured at zero licence cost
Commercial Savings	20% savings achieved
Commercial Structure	Pure token-consumption model
Commitment Risk	Significantly reduced
Governance	Full enterprise governance framework included
Ramp-up Approach	12-month adoption period

BUSINESS IMPACT

The most important outcome was not the discount.

The original structure required the client to commit to uncertain future demand before actual utilization could be measured.

The revised structure replaced fixed commitments with a consumption-based model, secured 500 enterprise seats at no licence cost, delivered immediate commercial savings of approximately 20%, and introduced a one-year ramp-up period supported by full enterprise governance controls.

The result was a significantly lower-risk commercial framework that preserved flexibility, improved cost transparency, and allowed the client to scale AI adoption as business demand matured.

KEY INSIGHT

In AI sourcing, some of the largest savings opportunities come from avoiding the wrong commitment before demand patterns are understood.

BY THE NUMBERS

PRICE VARIANCE

10x-50x

Observed by Chain IQ across comparable workload categories and commercial sourcing scenarios.

CASE STUDY SAVINGS

20%

Savings achieved through restructuring an enterprise AI agreement with Anthropic (Claude Code).

ENTERPRISE USERS

500

Enterprise users supported under the revised commercial framework.

PROCUREMENT PRIORITY

Flexibility over commitment

Commercial flexibility can deliver greater value than headline discounts in rapidly evolving AI markets.

TOKEN PRICING

Same job. Fifty times the price.

AI pricing varies significantly across providers, model tiers, deployment options, and negotiated commercial models.

> **CHAIN IQ BENCHMARK INSIGHT**
>
> The observed 10-50x variance is based on benchmarking of enterprise AI sourcing events, supplier proposals, client negotiations, public token pricing data, and commercial comparisons across multiple leading AI providers for comparable products/LLMs.
>
> Actual savings opportunities vary depending on workload complexity, supplier selection, deployment architecture, token consumption, and contract structure. The objective is not to identify the cheapest model. It is to identify the lowest-cost model that delivers the required business outcome and meets enterprise risk requirements.

Across comparable workloads, Chain IQ benchmark data and market observations regularly identify cost differences of approximately 10-50x between alternative model selections delivering acceptable business outcomes.

The important insight is not the absolute price. It is the variance.

For procurement leaders, every workload should first be assessed against the lowest-cost model capable of delivering the required outcome before being routed to more expensive reasoning models.

Using a premium reasoning model to perform routine administrative tasks is not a quality decision. **It is a cost leak.**

10-50x

The spread Chain IQ regularly sees between AI models delivering comparable business outcomes on the same workload.



CHAIN IQ'S POINT OF VIEW

Buy a portfolio – Not a platform

The goal is not replacing Microsoft, OpenAI, Anthropic, or Google. The objective is ensuring that no supplier becomes the only available option. Chain IQ increasingly views AI providers in tiers and applies four disciplines to manage the cost and risk that come with them.

01 See: build visibility before you scale

Most organizations buy AI before they can see how it is being consumed. The first move is a complete inventory of AI contracts, suppliers, and usage across the enterprise, down to workload level, so cost has an owner before it becomes a problem.

The objective is to establish a credible total cost and consumption baseline before entering supplier negotiations. Without this baseline, organizations risk negotiating attractive headline discounts while overlooking consumption growth, duplicate functionality, unused seats, expensive model defaults, or hidden infrastructure costs.

02 Tier: match the model to the workload and the user

Low-cost models, such as DeepSeek, Alibaba's Qwen, and other efficient alternatives, can absorb a large share of routine tasks. Mid-tier models suit customer-facing interactions such as simple chatbots. Premium reasoning models are reserved for genuinely complex work: legal analysis, financial modeling, high-stakes decisions.

Using a premium model for general administrative tasks, such as for a routine FAQ or document summary, by a back-office employee is not a quality choice. It is a cost leak.

However, model tiering cannot be based on price and technical performance alone. An economical model may not be acceptable for every organization, jurisdiction, data class, or use case. Each model must therefore pass a defined assessment covering security, privacy, residency, intellectual property, regulatory compliance, operational resilience, and geopolitical exposure.

MODEL APPROVAL FRAMEWORK

Cost should only become a selection criterion after minimum business and risk requirements have been satisfied.

Before approval, every model should be evaluated against:

- Business outcome requirements
- Accuracy and quality thresholds
- Data classification and sensitivity
- Security and privacy requirements
- Regulatory obligations
- Data residency requirements
- Operational resilience
- Total cost of ownership

Only after these criteria have been satisfied should organizations compare commercial economics and select the most cost-efficient option.

03 Portability has a cost

Many organizations believe portability is automatically valuable. In practice, portability requires governance, model testing, prompt adjustments, legal assessments, security reviews, data protection controls, and supplier risk management.

For regulated enterprises, the cost of introducing an additional supplier can exceed the immediate token savings.

The objective is therefore not maximum portability. The objective is to maintain credible alternatives where commercial leverage justifies the operational complexity.

04 Negotiate: convert competition into commercial value

Price negotiation remains a fundamental source of procurement value. The change is that AI negotiations require a broader and more dynamic commercial framework than conventional software agreements.

The starting point should be to separate the individual components of the supplier offer: enterprise seats, API consumption, token rates and discounts, implementation services, storage and cloud infrastructure. Combining these elements only makes benchmarking effective and levers attractive.

Negotiation priorities will differ by supplier:

Microsoft. With Microsoft, the strongest commercial leverage often comes from examining Copilot spend alongside Azure commitments, Microsoft 365 volumes, enterprise agreements, cloud consumption forecasts, and renewal timing. The discussion should extend far beyond the Copilot seat price. There will also be a considerable push from Microsoft to migrate from Copilot to Copilot Cowork, which introduces agentic functionality and token-based costing, and E7, which will bundle Copilot and additional AI features.

OpenAI. With OpenAI, the key negotiation point is separating enterprise seats from API consumption and avoiding premature long-term volume commitments while adoption patterns remain uncertain.

Anthropic. With Anthropic, successful negotiations frequently focus on commitment flexibility, tier migration rights, consumption credits, and phased growth assumptions rather than headline pricing alone.

Google. With Google, the most valuable commercial opportunities are often created by linking Gemini and Vertex AI discussions to broader cloud economics and existing hyperscaler commitments.

Across suppliers, the most important negotiation levers include volume discounts, token credits, free or discounted enterprise seats, phased commitments, price-protection clauses, indexation caps, service credits, model downgrade rights, audit rights, portability, termination assistance, and restrictions on using enterprise data for model training.

KEY INSIGHT

> The greatest value does not come from price negotiation or model selection in isolation. It comes from combining rigorous commercial negotiation, continuous supplier competition, and intelligent routing of each workload to the right model.

Ivo Claus Global Head of IT Sourcing, Chain IQ

What procurement leaders should take from this conversation

Consumption, not licensing

AI cost is reframed as a consumption-management challenge, governed at workload level rather than per seat.

A tiered supplier portfolio

Routing routine, mid-tier, and premium workloads to the right model cuts cost without cutting capability.

The portfolio must be aligned with an approved risk framework. Not every model will be suitable for every jurisdiction, data classification, or regulated environment.

AI FinOps in place

Consumption tracked at workload level, with finance, IT, and procurement aligned on ownership.

Procurement should use this consumption evidence in supplier negotiations, renewal decisions, model-routing policies, and demand-management discussions.

WHO OPERATES AI FINOPS?

Successful AI cost management requires coordinated ownership across multiple functions.

FUNCTION	PRIMARY RESPONSIBILITY
Business	Workload ownership and value realization
IT	Technical operations, model management, routing and monitoring
Finance	Budget ownership, forecasting and cost visibility
Procurement	Supplier governance, benchmarking and commercial optimization

Consumption should be measured at workload level and reviewed regularly against business value, model performance, security requirements, and commercial efficiency.

Without clear ownership and accountability, organizations often discover AI costs only after spend has accelerated.

10-50x price differences across comparable workloads

Organizations frequently discover significant cost differences between model providers and model tiers capable of delivering comparable business outcomes.

Continuous benchmarking

One-time negotiation is no longer sufficient. Procurement teams require continuous intelligence on supplier pricing trends, commercial practices, model capabilities, commitment structures, and emerging alternatives.

Chain IQ's negotiation experience and cross-client benchmarking activity provide a view of real commercial outcomes rather than public list pricing.

Governance ahead of cost

Frameworks built proactively, before overruns force a reactive scramble.

Residency can override headline price

An attractive model price does not automatically make a supplier suitable. Deployment location, jurisdiction, infrastructure, subprocessors, and political risk can determine whether a model is approved for a particular use case.

WHERE TEAMS GO WRONG

MISTAKE 01

Buying before seeing

AI is purchased before there is any visibility into how it will be consumed.

MISTAKE 02

Cost per token, not outcome

Teams measure spend per token instead of cost per business outcome delivered.

MISTAKE 03

Assuming pricing is stable

Today's pricing models will not last. New suppliers and capabilities arrive every few months.

MISTAKE 04

Negotiating every supplier in the same way

Microsoft, OpenAI, Anthropic, Google, API-first providers, and open-source solutions have different business models and therefore require different negotiation strategies.

MISTAKE 05

Failing to account for data residency requirements and cloud partner

Not every solution can achieve your data residency requirements. Some solutions will not be able to accommodate EMEA or specific European countries or will only be able to meet those requirements with certain public cloud providers, e.g. AWS. Check your data residency requirements against provider capabilities.

WHERE THE RISK REALLY SITS

The biggest procurement risk may not be paying too much

Historically procurement focused on achieving the lowest available unit price.

AI introduces a different strategic challenge. The market has consistently demonstrated falling prices and increasing capability.

The greater risk is no longer paying too much today. Organizations that lock themselves into large multi-year commitments may find themselves paying for capacity, pricing structures, or model capabilities that become materially less competitive within a relatively short period.

> The greater risk may be committing too much for tomorrow.

Joseph Sweet Global Head Lead Buying & Dynamic Pricing, Chain IQ

From market insight to executable AI sourcing

Chain IQ supports clients throughout the entire AI sourcing lifecycle, from supplier selection and benchmarking to contract negotiations and ongoing commercial governance.

Specialized AI Sourcing Expertise

Dedicated Chain IQ sourcing specialists focused on AI, software, cloud, and consumption-based commercial models.

Market Intelligence & Benchmarking

Access to benchmark data, pricing trends, negotiation insights, and commercial structures gained through continuous engagement with leading AI providers and enterprise clients.

Supplier Negotiation Experience

Hands-on negotiation experience with suppliers including Microsoft, OpenAI, Anthropic, Google, AWS, SAP, Oracle, Salesforce, and other leading technology providers.

Commercial Strategy & Risk Management

Support in structuring flexible agreements, challenging long-term commitments, securing commercial protections, and aligning AI sourcing decisions with regulatory, security, and business requirements.

A Broader Market Perspective

While most procurement teams see a limited number of AI contracts, Chain IQ leverages experience across multiple suppliers globally, industries, and sourcing events to help clients make better-informed decisions.

WHAT CHANGES WHEN CHAIN IQ IS INVOLVED?

Most procurement teams negotiate only a small number of AI agreements each year.

Chain IQ supports sourcing events, commercial negotiations, benchmark exercises, contract reviews, and supplier engagements across multiple clients, suppliers, industries, and geographies.

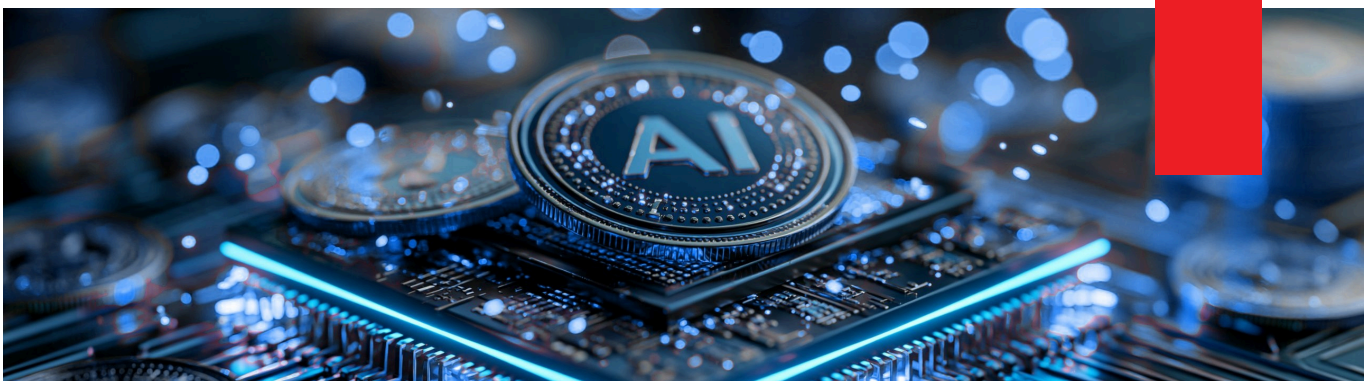
The difference is not procurement capability.

The difference is market visibility, negotiation experience, and access to benchmark intelligence derived from multiple sourcing events rather than a single enterprise agreement.

This provides clients with:

- Broader market visibility
- Real benchmark intelligence
- Proven negotiation strategies
- Enhanced commercial flexibility
- Reduced commitment risk
- Faster access to supplier and market insights
- Commercial structures tailored to rapidly evolving AI consumption models

Chain IQ helps clients turn AI sourcing from a technology purchasing exercise into a commercial advantage.



BUSINESS IMPACT

The cheapest model is not the answer

A competitive commercial agreement remains essential, but the most important decision is often not price. It is flexibility.

The organizations that succeed will not necessarily negotiate the lowest token price. They will design sourcing portfolios that create competition, preserve flexibility, maintain governance, and route each workload to the most economically appropriate model.

Because in AI, the most valuable concession is often not a discount. It is the right to change your mind.

> The organizations that win won't be the ones with the cheapest AI supplier. They'll be the ones that treat AI sourcing as a portfolio, and keep their architecture open enough to keep choosing.
>
> **Joseph Sweet** Global Head Lead Buying & Dynamic Pricing, Chain IQ

WHERE TO START

01 · TOMORROW MORNING

Build the inventory

A complete map of AI contracts, suppliers, and consumption across the enterprise.

02 · THIS QUARTER

Introduce AI FinOps

Track consumption at workload level; give cost an owner before it scales.

03 · THIS QUARTER

Establish an AI supplier and residency policy

Define approved providers by workload, data classification, processing region, jurisdiction, regulatory requirement, and geopolitical risk profile.

04 · AT THE NEXT SOURCING EVENT OR RENEWAL

Negotiate the complete commercial model

Separate seats from consumption; challenge volume commitments; negotiate token credits, caching and batch discounts, ramp-up structures, price protection, service credits, model flexibility, data protections, and exit rights.

05 · ONGOING

Benchmark continuously

Create competition between suppliers, for example via pricepertoken.com. Reassess price, capability, residency, risk, and contractual maturity as suppliers release new models and commercial structures.

FEATURED EXPERTS



Ivo Claus
Global Head of IT Sourcing, Chain IQ



Joe Sweet
Global Head of Lead Buying, Chain IQ

© 2026 Chain IQ Group AG · www.chainiq.com · Privacy