OFFICE HOURS REPORT — NO. 01

Managing AI costs

Rethink how ~~much~~ you charge.

DATE

May 21, 2026

FORMAT

Online, live Q&A

REGISTERED

47 practitioners

QUESTIONS

30+ in advance

Ulrik Lehrskov-Schmidt

ANSWERING LIVE — MAY 21, 2026

This report is an AI assisted product made with the collected knowledge of Willingness to Pay and validated by a human.

Contents

EXECUTIVE SUMMARY

The short version

AI broke the fixed-cost economics that SaaS pricing was built on: every query now carries a real cost, and flat-fee unlimited plans quietly hand your margin to your heaviest users. This report answers the questions 47 pricing practitioners brought to our May 2026 Office Hours: how to price AI products when costs are unknown, how to square variable costs with customers who demand predictable bills, what good fair use policies look like, and when outcome-based pricing works. The answer runs through three principles: see your usage, set explicit limits, and price the value instead of the compute. Cost sets your floor. Value sets your price. The structure in between is where you make your money.

01 — THE STARTING POINT

AI broke the margin model. That is fine, as long as you know it.

Traditional SaaS economics are a beautiful thing. You build the product once, you host it, and every additional customer costs you close to nothing. That is how the industry got to 80% gross margins, and it is why nobody had to think very hard about cost when designing a pricing model. Cost was a rounding error.

AI ends that. Every query, every API call, every token has a real cost attached, and the more your customers use the feature, the more it costs you to serve them. If you price AI like traditional SaaS (flat fee, unlimited usage) you have built an adverse selection machine: your heaviest users crush your margins and your lightest users subsidize them. This is not theoretical. OpenAI reportedly struggled with heavy users on its top-tier plans, and Atlassian publicly tied a 10% price increase directly to rising AI infrastructure costs.

At the same time, the ground keeps moving under your feet. Token prices dropped roughly 97% through 2024, and then better models invited bigger tasks, so consumption went up while unit costs came down. Whatever AI can do today, it can probably do twice as well tomorrow at half the cost, and nobody knows the exact rate. So don't try to precision-engineer something that can't be precision-engineered. Build a structure that absorbs the movement instead.

Roughly half of our current consulting projects have large AI components, and what we do in every single one of them comes down to three principles:

1

See your usage

You cannot manage a cost you cannot see. Instrument every AI feature. Know which customers use what, and where cost accumulates by segment, feature and use case. Granular enough to act on.

2

Set explicit limits

Unlimited usage is great marketing copy and a margin disaster in practice. Included volumes, fair usage policies, soft and hard limits. Cap your downside before a customer runs a thousand queries an hour.

3

Price the value, not the compute

Do not pass your infrastructure bill through to the customer. Price what the outcome is worth. Then falling AI costs become your margin instead of your customer's discount.

> **Cost sets your floor. Value sets your price. The structure in between is where you make your money.**

02 — YOUR QUESTIONS

What you asked, and what I would do

We grouped the questions you submitted into four themes. Where several of you asked variations of the same thing, we merged them. The answers are the short version of what I do with clients; your business could defensibly land somewhere different, but the direction would be the same.

Launching when costs are unknown

? How do we price a new AI product when we can't predict customer usage or our own costs, and the costs are probably unoptimized anyway?

You run two models at once. For 80–90% of customers, price on a structure your CFO will accept: an included volume, a conservative cost-anchored price, guardrails. Then take a petri dish (say 5 of your 500 customers) and give them all-you-can-eat for a fixed fee. One of two things happens. Either consumption stays low, which tells you that you have a product problem and not a pricing problem. Or consumption goes high, and you either find cost without value (product problem again) or your best value cases, which now tell you exactly what to charge everyone else. Either way you have de-risked the model. Don't wait for perfect data; design for learning.

? How easy is it to estimate future compute cost?

Harder than your spreadsheet thinks. Unit costs collapsed (roughly 97% on tokens through 2024) while better models invited heavier tasks, so total consumption rose anyway. Both forces will continue and nobody knows the rates. I would not build a 3–5 year cost forecast; that is a category error. Build a cost curve you re-read every quarter, and a pricing structure that doesn't need the forecast to be right.

? Where are the chunky bits in AI cost, and how do I align cost and revenue tighter?

You find the chunky bits empirically, not theoretically: instrument first, then look at cost by customer, by feature and by use case. In practice a small share of customers and one or two features usually carry most of the cost. Once you can see that, alignment is a structure question: move the heavy use cases onto a metered or credit-based component and leave the light ones in the base. If you cannot explain your cost curve by segment, you are not ready to scale the product.

Predictability versus variable cost

? Customers want predictable bills. Our costs scale with usage. How do we square that?

Three tools: measurability, predictability, controls. Make the metered thing trackable in real time, forecast it into money for the customer ("at this run-rate you hit your limit in 4 months, which means another $12,000"), and give admins caps and permissions so they control who spends. And then consider the credit system, which is usage-based pricing paid up front: the customer budgets one number, you get the cash, unused credits roll over. The customer experiences a fixed spend; you keep the usage economics. That tension between variable cost and predictable billing is not a paradox to solve, it is a structure to build.

> **"Procurement does not hate variability, it hates surprises."**
>
> ULRIK LEHRSKOV-SCHMIDT — OFFICE HOURS, MAY 2026

? Will enterprise procurement ever accept variable agreements based on LLM usage?

Procurement does not hate variability, it hates surprises. What I see getting through purchasing departments is the commit: a credit pool sized together with the customer, paid up front, with terms that guarantee they never overpay. No overage penalty (top-up credits cost the same), and apply volume discounts retrospectively so under-committing is never punished. Psychologically the customer is depositing money, not spending it, which is an easier yes than a license commit of the same size. Start with a handful of accounts, not a big-bang migration.

? Tokens or credits? Users get tokens, but procurement struggles with them.

Tokens are infrastructure pricing. If you sell raw horizontal capability the way OpenAI does, fine. If you sell a solution, sell credits: an abstraction layer between dollars and your product. Keep the dollar-to-credit ratio dead simple (1:1 or 1:10, log-scale, no weird math) and put all the complexity between credits and units instead, where you can tune it. A buyer can budget credits. Nobody budgets tokens.

? We baseline normal usage, cap it at $x per month, and let customers buy more. Reasonable?

Reasonable, with one change: don't punish the top-up. When they run out and buy more, the price per unit should not go up, and if the larger total volume would have earned a discount, apply it retroactively. The moment a customer expands is the worst possible moment to add friction. You want the cap to start conversations, not stop usage.

? We burn credits by query complexity buckets (low, medium, high). Pros and cons?

It's a sensible variance-management move, the same trick as charging API calls by the second instead of by a denser metric. Three cautions. Publish the buckets and how you sort into them in real time, or it becomes a black box and black boxes erode trust. Expect to re-draw the boundaries as your costs fall, so don't present them as permanent. And treat it as a learning model: once you know which use cases carry the value, migrate pricing from query complexity (an input) toward the outcome the queries produce. Input pricing is where you start, not where you finish.

Fair use, spikes and over-usage

? What do good fair use policies and limits actually look like?

A fair usage policy does three different jobs, and you should know which one you are hiring it for: a monetization threshold (hit the limit, pay more), a variance control (cap overuse), and an oversell guard (a soft limit that triggers notifications to the customer and your CS team, and a hard limit before it threatens the platform). Set the soft limit where 95–98% of customers never touch it. Their legal team will find the clause; that is the right place for that discussion. And enforce it: I worked with a European scale-up (roughly €30M ARR, 6,000 customers) where the unenforced fair usage clause had quietly accumulated the legal right to invoice an extra €28M in 12 months. We didn't send that invoice, but it told us exactly which metric the next pricing model should be built on.

"Unlimited usage is great marketing copy and a margin disaster in practice."

ULRIK LEHRSKOV-SCHMIDT — OFFICE HOURS, MAY 2026

? A customer hits a usage spike. What signals should we act on before throttling, without penalizing high-value customers who are just using the product as intended?

First separate value-generating usage from cost-generating usage, because they need opposite responses. In my experience the unfair usage is concentrated in fewer than 5% of customers, and some of them run 10x to 200x the policy. You cannot invoice 200x retroactively without ever having had the commercial conversation, so don't try; reach out, apologize for not having had the right limits in place, and restructure their plan. And the heavy user who is simply succeeding with your product is not a breach, they are your expansion pipeline: the soft limit should trigger a customer success conversation, not a punishment. Reserve hard stops and throttling for the cases where the spike threatens the platform itself.

Value, outcomes and margins

? Some AI features cost us almost nothing but deliver enormous value. How do we bridge that disconnect?

There is no disconnect to bridge, because cost and value were never supposed to correlate. Cost sets the floor below which you refuse to sell. Willingness to pay sets the price. The gap between them is your margin, and with AI that gap just varies wildly from feature to feature. Practically: price the high-value features on the value they create (fence them into the right tier or package), and let the high-cost features carry metering so the floor is protected. Price the customer, not the product.

? Outcome-based pricing is alluring for AI. Does it apply to every vertical and use case?

No, and I would not force it. The question is whether you are horizontal or vertical. Horizontal products (the AI can do anything for anyone) cannot credibly price outcomes, because they don't own any specific outcome; they price inputs, like tokens, and they compete on volume. Vertical products that own a measurable, attributable outcome in a value chain the customer already prices (a resolved ticket, a collected payment) can and should price the outcome. Business units with a flat knowledge-worker cost base and no output measure fail the attribution test, so charge them for the job to be done instead. Outcome pricing is just value pricing with better instrumentation; it existed before AI and the rules did not change.

? What is the minimum margin we should accept on an AI product?

There is no universal number, and in the launch phase it is the wrong question. The right question is whether you can explain your AI cost curve by segment, feature and use case. If you can't, don't scale, because every new customer might be making you poorer. Once you can, I would expect software-level gross margins to return over time as compute costs fall, provided your pricing is anchored to value rather than cost-plus. If you contracted your prices to the compute bill, the falling costs belong to your customer. If you priced the outcome, they belong to you.

? How do we combine usage-based pricing with other metrics, like per user? And how do we keep the model simple enough to understand?

Combining metrics is normal, not exotic: a flat platform fee for the things everyone needs, license metrics (users, locations) for the predictable capacity, and a consumption or credit metric for the variable AI workload. One metric carries the value story; the others cover cost and structure. On simplicity: customers don't actually need a simple model, they need an explainable one. They will accept any metric they can measure, predict and control, and they will accept it twice as fast if they already pay for it somewhere else. Spend your complexity where the customer sees value, and bury the rest in the terms.

? Our industry is used to fixed subscriptions. When should we move first toward usage-based or AI-cost-sensitive pricing, and when should we stay put?

Habit is one of the strongest fairness drivers in pricing: buyers accept metrics they already pay for elsewhere, and resist ones they don't. So don't be the missionary unless your cost exposure forces you to be. The good news is you rarely face a binary choice. A credit commit looks like a subscription to procurement (one number, paid up front) and behaves like usage inside the product. That is usually how I take a conservative market across the bridge: keep the buying experience familiar, change the engine underneath.

? How do we change pricing after customers have already built usage of our AI product?

This is playing out in public right now: most of the companies that launched AI as a $4–$10 add-on (Notion, Slack, Loom) have folded it back into the core and re-anchored the base price. That is the normal three-phase arc. Phase one, the add-on experiment: test willingness to pay without touching the core. Phase two, data collection: learn usage intensity and cost exposure by segment. Phase three, rebundling: AI becomes table stakes, goes into the base, and the base price moves up. Most companies stall between two and three because rebundling requires repricing courage. If your customers now expect AI by default, the question is no longer what to charge for the add-on. It is what your core product is worth now.

03 — YOUR NEXT STEPS

What to do in the next 90 days

Not a transformation program, just the sequence I would run with your team. Each step makes the next one cheaper.

1 **Instrument before you price**

Get usage and cost data per customer, per feature, per use case. Every decision below depends on this, and most teams discover their assumptions about heavy users were wrong.

2 **Draw your cost curve**

Find the chunky bits. Which 5% of customers and which one or two features carry most of the cost? Can you explain the curve to your CFO in one slide? If not, repeat step 1.

3 **Pick your bet: value or commodity**

Decide whether you are building toward owning specific outcomes (vertical, price the output) or providing broad capability (horizontal, price the input). How you price today follows from where the product is going, not from this quarter's compute bill.

4 **Choose your modality**

Flat fee, license, usage or credits. For B2B AI workloads with variable consumption, credits are usually the strongest default: predictable for the buyer, covered costs for you, faster adoption than pure usage. Keep the dollar-to-credit math dead simple.

5 **Set the guardrails**

Included volumes, a soft limit that alerts the customer and your CS team, a hard limit that protects the platform. Place the soft limit where 95–98% of customers never touch it. No overage penalties; same unit price on top-ups, discounts applied retroactively.

6 **Run the petri dish**

Give a small subset of customers unlimited usage at a fixed fee and watch what happens without pricing friction. Use what you learn to set the model for everyone else, then revisit quarterly, because your costs will move again.

04 — ONE MAP TO KEEP

Where does your AI product sit?

Most of the questions in this session resolve differently depending on one structural choice: whether your AI is horizontal or vertical. Before your next pricing discussion, agree internally on which row you are in.

POSITION	WHAT YOU SELL	WHAT TO PRICE	COST MANAGEMENT
Horizontal	Broad capability; the AI can do almost anything for almost anyone	Inputs: tokens, requests, compute, credits as utility	Pass-through economics with thin margin; win on volume and efficiency; metering is the product
Vertical	A specific outcome in a value chain you know and the customer already budgets for	Outputs and outcomes: resolved tickets, processed payslips, completed jobs	Cost is your floor, not your price; guardrails protect margin while falling compute costs expand it

If you are horizontal today but vertical tomorrow, price for learning now (credits, simple metrics, petri-dish experiments) and move to outcomes as soon as you lock in a value chain. The expensive mistake is staying on input pricing after you own the outcome.

KEEP LEARNING — FROM THE SESSION ARCHIVE

Agentic AI Pricing webinar series (6 parts) — strategy, packaging, models, tokens and credits, expansion, case studies

AI Features and the Margin Trap — the three principles of margin-safe AI monetization in 3 minutes

Why SaaS Companies Are Quietly Killing AI Add-Ons — the three-phase arc from add-on to rebundling

Fair Usage Policies Explained — thresholds, variance control and soft/hard limits in practice

The Pricing Roadmap — the book behind the frameworks

WORK WITH US

Rethink how much you charge.

We design AI pricing models end-to-end: strategy, packaging, metrics, guardrails and implementation. We specialize in transformation, not optimization, and roughly half of our current projects have large AI components. We can't help you tinker with your pricing. But if you're ready for a redesign, connect with us.

200+
PRICING REDESIGNS

242%
AVERAGE UPLIFT

125
AVG. DAYS TO LAUNCH

0
FAILURES OR BLOW-UPS

Schedule a Call

or email ulrik@willingnesstopay.com

SENIOR PRICING ADVISORS

Ulrik Lehrskov-Schmidt

CEO, FOUNDER

A globally recognized authority on pricing strategy and author of The Pricing Roadmap. Trusted advisor to global SaaS companies for over 20 years.

Christopher Truce

COO & CO-FOUNDER

A seasoned expert in financial services and SaaS products with 18+ years across global markets, from product development to commercial strategy and pricing innovation.

Roee Hartuv

SENIOR PRICING ADVISOR & HEAD OF GTM

A B2B SaaS executive with over 20 years driving revenue growth at high-growth companies, from sales and customer success to the consulting practice at Winning by Design.

Morten Klank

SENIOR PRICING ADVISOR

A seasoned SaaS executive with over 20 years of leadership experience in complex change, strategic repositioning and several high-impact turnarounds in PE-backed companies.

This report is an AI assisted product made with the collected knowledge of Willingness to Pay and validated by a human.