OWNED INTELLIGENCE, EXPLAINED

The Economics of *Owned Intelligence*

Why grocery's AI bill doesn't have to scale with its business, and what happens to the cost curve once a retailer owns the intelligence instead of renting it.

Joseph Lee

CEO & Co-Founder, Delectable AI

14x

Lower cost per AI session vs. a general-purpose LLM baseline

490ms

Household-aware response latency, roughly 8x faster than a cold-context baseline

$3.8M

Projected annual cost avoidance at ~125M sessions/yr

Zero

Off-catalog recommendations across thousands of pilot sessions

Every grocery executive I sit across the table from this year asks some version of the same question. Not "should we do AI." That debate ended sometime around when Walmart put Sparky inside ChatGPT and Kroger signed a national partnership with Google Cloud and Gemini. The question now is

narrower, and harder: what does it actually cost to run this at scale, and who controls that cost curve?

That question deserves a real answer, with real numbers, not a vendor's assurance that "it'll get cheaper." So I want to walk through the actual economics. Where the savings come from, why they compound rather than erode, and why I think the next eighteen months will separate grocers who rent their intelligence from grocers who own it.

01 · THE HIDDEN COST CURVE

The hidden invoice

Start with the uncomfortable part. If you build a grocery AI experience directly on top of a general-purpose LLM, which is the path Walmart took with OpenAI and the path Kroger is now taking with Gemini, you are not buying a fixed-cost platform. You are buying a meter. Every prompt, every reasoning step, every retry runs a token bill, and that bill scales linearly at best with your shopper adoption. The better your AI experience performs, the more it costs you to run it. That's an inverted incentive for any retailer with a P&L to defend.

We instrumented this directly rather than take anyone's word for it, our own included. In our Q1 2026 pilot with Giant Eagle, we logged the full token stream, prompt and completion, on both stacks in parallel, against the same shopper query, the same catalog scope, and the same response constraints.

Thousands	$0.033 → $0.0024	14x	$3.8M
AI-influenced sessions instrumented head-to-head in our Q1 2026 pilot with Giant Eagle	Average cost per session: general-purpose LLM baseline compared with Delectable's architecture	Cost ratio between the two stacks, statistically significant across the pilot	Modeled annual cost avoidance at ~125M sessions in a Year 3 regional-grocer projection

Fourteen times isn't a rounding error. At the scale a national or large regional grocer runs, meaning tens of millions of AI-influenced sessions a year, growing every quarter as adoption climbs, the gap between those two cost curves is the gap between an AI program your CFO can underwrite with confidence and one that becomes a standing line-item risk on every board deck. We modeled this out to Year 3 for one regional partner at a

projected ~125 million annual AI sessions. The cost avoidance versus a rented-LLM baseline lands around **$3.8M a year**. That's not revenue we're claiming credit for. That's the invoice that never gets sent.

02 · WHY IT'S STRUCTURAL

Why the discount is *structural*, not promotional

Here's the part that I think gets missed in most AI vendor pitches. Our cost advantage isn't a subsidized rate we're offering to win your business. It's architectural, and it comes from where the intelligence is actually being sourced.

A general-purpose model has to reason over a massive store of data every time it's asked a grocery question. It doesn't know what's on your shelf today, what your shopper bought last Tuesday, or what's actually in stock at the store she's walking into. So the burden of constraining, grounding, and personalizing that answer gets pushed into the prompt, which means more tokens, more reasoning steps, more retries, and a fatter, more unpredictable cost distribution.

Delectable inverts that order of operations. We build two proprietary knowledge graphs, seeded by your own first-party data.

OWNED DATA LAYER · 01

The Food HyperGraph

Encodes your own catalog at up to 42 attributes per SKU (ingredient, flavor, nutrition, allergen, cuisine, meal pattern) so retrieval starts from ground truth rather than an open corpus.

OWNED DATA LAYER · 02

The Shopper HyperGraph

Learns household preference, budget shape, and intent from your own first-party signal, seeded by your existing loyalty account. Not a generic profile inferred mid-prompt.

Retrieval, ranking, filtering, and cart assembly all run deterministically against those graphs, built from your data, running on your tenancy. The general-purpose LLM only gets called at the very end, for final language formatting, after the data has already been filtered down to what's actually true, in stock, and relevant to that household. In our architecture that final decision, fusing both HyperGraphs into a single household-aware

answer, resolves in **roughly 490 milliseconds**, about 8x faster than a ~4.0-second baseline from a general-purpose model doing the reasoning itself.

> *"Cheaper, faster, and, because every response is filtered against your verified SKU catalog before it ever reaches a shopper, architecturally incapable of hallucinating a product you don't sell.*

That's the mechanism, plainly stated. Source the intelligence from data you already own, constrain the output to inventory that's actually real, and call the expensive general-purpose model only for the narrow job it's actually needed for. In our pilot instrumentation, that came out to **zero off-catalog recommendations** across thousands of sessions. Not a low rate. Zero.

03 · WHERE THE CURVE IS HEADED

From rented tokens to *owned* silicon.

Enterprises don't rent CPU cycles forever, not once a workload is large enough and predictable enough to justify owning the infrastructure. Grocery AI, run at the transaction volume of a 150-to-1,500-store chain, crosses that threshold faster than almost any other AI use case in retail. The query pattern is narrow, even as the volume is enormous.

I'd be doing this topic a disservice if I stopped at "we're cheaper today" without addressing where the industry's headed from an infrastructure and token-cost perspective, because I think we're at the cusp of a second shift that most retail AI conversations haven't caught up to yet.

The first generation of grocery AI, the Walmart-OpenAI, Kroger-Gemini, Instacart-everyone pattern, treats intelligence as a utility you meter and pay for per call, hosted on someone else's cloud, trained in part on someone else's aggregate data. That was a reasonable starting point when foundation models were the only path to a credible shopping assistant. It is increasingly not the only path, and for high-volume, narrow-domain applications like grocery, it's not even the most economical one.

What I'm watching now, and what we're building toward with our own architecture, is the migration of the deterministic, high-frequency parts of the reasoning stack onto smaller, fine-tuned, increasingly locally-hosted or owned-hardware models, with general-purpose frontier LLMs called only for the residual cases that genuinely require open-domain reasoning. This isn't a contrarian bet; it's the same maturation curve every compute-intensive industry has gone through. Enterprises don't rent CPU cycles forever once the workload is large enough and predictable enough to justify owning the infrastructure. They build for it. Grocery AI, at the transaction volume of a 150-to-1,500-store chain, crosses that threshold faster than almost any other AI use case in retail, because the query pattern is narrow (a bounded catalog, a bounded set of households, a bounded set of intents) even as the volume is enormous.

The economic logic is straightforward once you see it. Every token you send to a rented, general-purpose model is a marginal cost that scales with your success. Every token you can instead resolve on a smaller model you've fine-tuned on your own catalog and household data, whether that's hosted on your own cloud tenancy today or, over the medium term, on owned or reserved hardware, is a cost that starts to look more like depreciation than a utility bill. That's the direction the unit economics point, and it's why we architected Delectable from day one to be deterministic-first, with LLM calls as the exception rather than the rule. Retailers who build their AI program assuming the token meter is a permanent fact of life are underwriting a cost structure that only gets more expensive as they succeed. Retailers who build toward owned, narrow, fine-tuned models on top of their own first-party data are underwriting a cost structure that gets cheaper as they scale. **That gap compounds every fiscal year.**

What this means for the *consumer's wallet*

Cost efficiency on the backend only matters if it translates into value the shopper actually feels, so let me turn to the demand side, because the savings there are just as real and, I'd argue, even more durable.

BUDGET CONTROL, BUILT IN

AI-powered budgeting inside the cart itself

The single biggest complaint we hear in consumer research isn't about selection. It's about the gap between what a household intends to spend and what actually shows up on the receipt. A household-aware agent that knows the budget going in ("$150 for the week," "SNAP-eligible items only," "keep it under $40 for the birthday dinner") and constrains every recommendation to that ceiling in real time does something a generic search box structurally cannot. It prevents the overspend before it happens, rather than surfacing a coupon after the fact. That's a fundamentally different economic relationship with the shopper than "here's 40% off if you clip this digital coupon." It's proactive cost control, not reactive discounting.

PERFECT CART

A pantry-aware meal planner is a food-waste engine wearing a convenience feature's clothes

The average American household throws away a meaningful share of the groceries it buys, and most meal-planning tools make that worse, not better, because they plan from a recipe outward rather than from the refrigerator inward. A system that remembers what's already in the pantry (what's been purchased, what hasn't been used yet, what's about to expire) and builds the week's meal plan to use that inventory first before recommending net-new purchases is doing real household economics, not just recipe suggestion. It's the difference between a system that maximizes basket size and one that earns trust by minimizing waste. We've built our Perfect Cart experience around exactly that principle. It learns the household, remembers the pantry, and crafts each cart dynamically against live inventory, not a static shopping list.

Both of these, budgeting and pantry-awareness, are also, not coincidentally, the features that build the kind of trust that keeps a household loyal for years rather than one promotional cycle. Consumer economics and retailer economics point the same direction here, which is rare enough in retail AI that it's worth calling out directly.

What this means for the *grocer's cost structure*

Now to the side of the ledger that, frankly, gets less attention in most AI vendor conversations but matters just as much to a CFO. The operating cost reductions inside the retailer's own four walls.

Customer service load drops when the shopping experience stops generating the questions in the first place. A meaningful share of grocery e-commerce customer service volume traces back to friction in discovery and substitution. The shopper couldn't find what she wanted, or an out-of-stock item got swapped for something that violated a dietary restriction or a brand preference she'd never have accepted. An agentic layer that reasons over allergen, dietary, and preference constraints before the substitution is ever offered, rather than after a complaint is filed, removes that friction at the source. Fewer contacts, fewer refunds, fewer escalations, and a materially better NPS on the exact moment (the out-of-stock swap) that most reliably erodes trust in digital grocery.

01

Retail media, automated.

Running a modern RMN well requires campaign planning, inventory targeting, creative rotation, reporting, and reconciliation across dozens of CPG partners at once. In our modeling with one regional retail-media partner, agentic workflow automation (intelligent campaign planning, automated approval routing, CPG self-service tooling) collapses what has historically required a five-person operations team down to **one or two FTEs**, while improving CPG reporting fidelity because every recommendation now traces to a specific household signal rather than a historical look-back file.

02

Revenue stays on your rail.

That same layer keeps the CPG relationship, and the associated revenue, running through the retailer's own owned-media rail instead of a third-party network's. Our pilot data shows retailers retaining roughly **95% of recaptured retail-media revenue** net of platform costs, rather than ceding the CPG dollar to a middleman's agency or a third-party RMN.

03

Acquisition cost, reinvented.

Delectable Social collects engagement metadata from shoppable, creator-led video content and uses it to build an actual community layer around the retailer's brand. Regional creators, culturally specific meal content, shoppable recipes tied directly to what's in stock at the shopper's home store. The result is **lower CAC and higher LTV from the same channel**, because the content, the commerce, and the household memory are running on one intelligence layer instead of three disconnected systems.

And then there's acquisition cost, the number that quietly determines whether a loyalty program is an asset or a subsidy. Grocery has historically been a poor acquisition channel for Gen Z and Millennial households, who don't respond to a Sunday circular and increasingly don't respond to generic social advertising either. Lower CAC and higher LTV from the same channel isn't the kind of result that happens by accident. It happens because the content, the commerce, and the household memory are all running on one intelligence layer instead of three disconnected systems.

06 · THE THREAD CONNECTING ALL OF IT

Three moves, *one* architecture.

If there's one idea I want an executive team to leave this article with, it's this. The cheapest AI architecture and the most trustworthy one are the same architecture. Everything else follows from three decisions.

01

Own the data.

Ground the system in a retailer's own catalog and a retailer's own household data, not an open corpus that has to be constrained on every single call.

02

Constrain the model.

Filter every response against the verified SKU catalog before a shopper ever sees it. Cheaper to run, faster to respond, and structurally unable to hallucinate.

03

Call the expensive reasoning only when it's genuinely needed.

Reserve the general-purpose LLM for the narrow job of final language formatting, after retrieval, ranking, and filtering have already run deterministically.

That's not a coincidence, and it's not a coincidence I take lightly, because it would have been very easy for us to build this company the other way. Wrap a general-purpose model in a grocery-flavored prompt, ship fast, and let the token bill be tomorrow's problem. We didn't, because a system that's grounded in a retailer's own catalog and a retailer's own household data is, by construction, cheaper to run, faster to respond, harder to hallucinate with, and easier for a CFO and a CIO to both sign off on in the same room. Own the data. Constrain the model. Call the expensive reasoning only when it's genuinely needed. Everything else, the 14x cost advantage, the sub-second latency, the zero off-catalog recommendations, the compounding household intelligence that gets sharper with every basket, follows from that one architectural decision.

Grocery is the last major consumer category to make the shift to **AI-native discovery.**

The retailers who move first won't just win on shopper experience. They'll be the ones who built their AI program on a cost curve that bends down as they scale, while their competitors are still discovering what their token bill looks like at ten times today's volume.

Joseph Lee

CEO & Co-Founder, Delectable AI

Joseph Lee is CEO and Co-Founder of Delectable AI. Pilot figures referenced in this article are drawn from Delectable AI's Q1 2026 deployment with regional grocery partners. Full methodology and reference architecture are available under NDA.