The "Juice Isn't Worth the Squeeze"

How Architecture-Led AI Cracked a Pricing Model That Defeated a Two-Year Modernization Effort

A case study in applying AI and machine learning to one of the hardest problems in specialty insurance underwriting: pricing complex, high-cost cases when the data is sparse, the stakes are in the millions, and the legacy model is decades old.

Author: **Mitch Quinn, Director of AI/ML**

The Problem No One Had Solved

Imagine being asked to predict, with confidence, what a complex medical case will cost over the next several years. Now imagine that prediction is the basis for a multi-million-dollar fixed-price contract. Get it right and you protect both margin and patient outcomes. Get it wrong and you either leave money on the table or, worse, go underwater on a contract you cannot exit.

This is the daily reality for organizations that take on risk-based outcome plans for severe, high-acuity medical cases. Each case is unique. Each carries millions of dollars in potential cost. And the pricing decision must be made as quickly as possible to secure the contract and begin care management of the patient.

Our client, a leader in this space, has been using a model built decades ago to assist with pricing these contracts. The model assigned each patient a score from one to six based on roughly two dozen clinical variables, with weights derived manually from limited historical data. Cases are then labeled as low, medium, or highly complex in part based on this model prediction. It worked, sort of. But three problems had grown impossible to ignore.

First, complexity scores were overloaded. The same number meant different things to different stakeholders. To a clinician, complexity meant treatment intensity. To an underwriter, it meant uncertainty. Scores got vetoed in committee, debated, and re-derived case by case. The metric had drifted from being a business signal to being a starting point for argument.

> **Get it right and you protect both margin and patient outcomes. Get it wrong and you either leave money on the table or, worse, go underwater on a contract you cannot exit.**

Second, the model could not separate medium-risk cases from high-risk cases. There was significant overlap, with mediums that should have been highs and highs that were not high enough. That ambiguity sat directly on top of the company's profit margin.

Third, every prior attempt to modernize the calculator had failed. A two-year effort by a contracted data scientist added dozens of clinical variables but delivered no meaningful accuracy improvement; leading the client to conclude the operational burden of capturing that additional data element wasn't worth it. The juice was not worth the squeeze. The project was abandoned.

That is the door we walked through.

Why Architecture Came First

When most consultancies are asked to “do AI” on a problem like this, they reach for the model first. They pick an algorithm, point it at the data, and tune for accuracy. We did the opposite. Before we trained anything, we mapped the architecture of the entire decision-making process the calculator lived inside.

> **When most consultancies are asked to ‘do AI’ on a problem like this, they reach for the model first...We did the opposite.**

That mattered for a specific reason: this was not a greenfield AI project. It was a system already woven into a regulated underwriting workflow that touched clinical services associates, directors of clinical services, medical directors, an actuarial team, and a case review meeting where humans debated and overrode model outputs by design. Any AI we built had to land inside that workflow without breaking it.

Our healthcare and insurance industry experience told us three things up front that would have been easy to miss otherwise:

The data was going to be sparse. In specialty lines like this one, the cases that drive the economics are, by their nature, rare. Across nearly a decade of usable history (the cutoff being a process change that invalidated older records), we had roughly 2,500 complete cases to work with. For a regression model trying to predict a multi-year cost trajectory, that is far fewer cases than what a typical training set would contain.

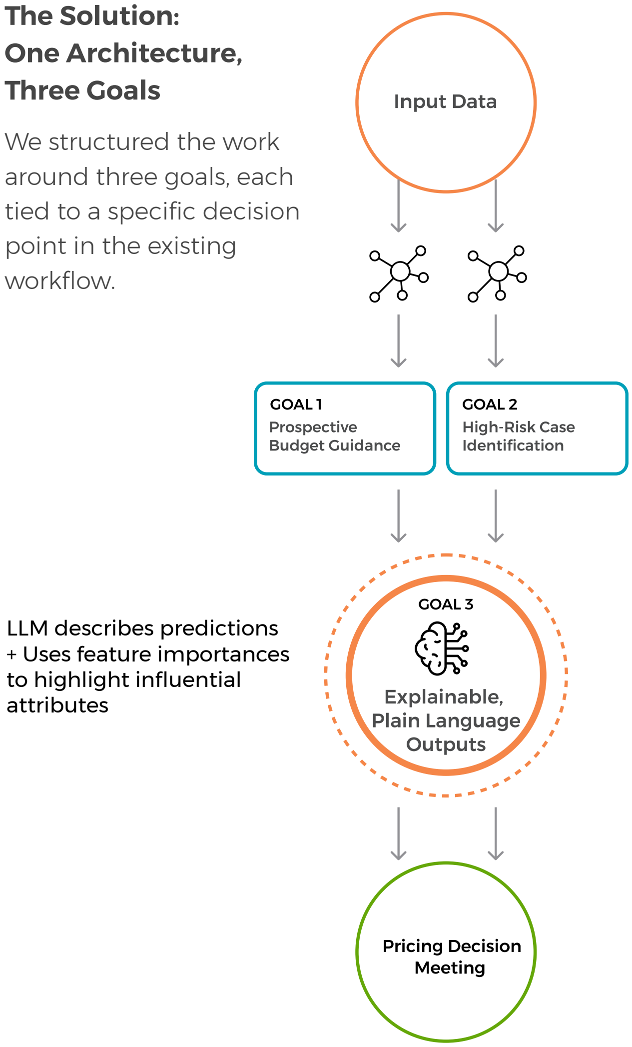
The output had to be interpretable. In healthcare and insurance, a black-box prediction is a non-starter. Underwriters, actuaries, and clinical directors will not, and should not, accept “the model said so” as the basis for a multi-million-dollar contract. Whatever we built had to surface the clinical and demographic factors driving each prediction in language a non-technical reviewer could defend in a meeting.

The output had to be integrated with how the business already worked. The calculator did not exist in isolation. It fed budget templates that were built bottom-up by clinicians, line by line, across roughly ten cost categories ranging from inpatient care to in-home services. Replacing that process was not on the table. Augmenting it, with high-confidence predictions that clinicians could compare against their own bottom-up estimates, was.

Once we had that architectural picture, the modeling decisions almost made themselves.

The Solution: One Architecture, Three Goals

We structured the work around three goals, each tied to a specific decision point in the existing workflow.



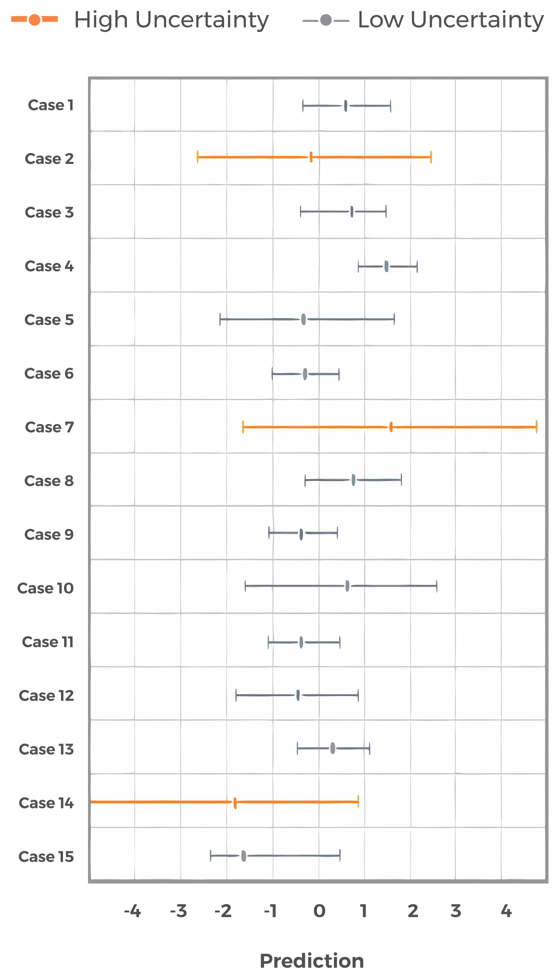
Goal one: prospective budget guidance.

Rather than build one monolithic model, we built ten. Each model predicts cost for one of the client’s existing summary categories: inpatient spend, outpatient spend, in-home care, and so on. Predictions are generated by category, then aggregated. State-level features account for regulatory and pricing variation. Historical costs were inflation-adjusted to current-year dollars before training, since healthcare CPI had moved dramatically across the historical window.

The result: when we benchmark our predictions against the client’s existing low, medium, and high complexity tiers using actual historical spend, our model produces tighter, better-separated cost ranges. Our “low” predictions are lower than their lows. Our “high” predictions are higher than their highs. The medium-versus-high overlap problem that had plagued the legacy model is materially reduced. The client’s own assessment, in their words, is that even if you boiled our outputs back down to a one-through-six scale, the result would already be an improvement over decades of incumbent methodology.

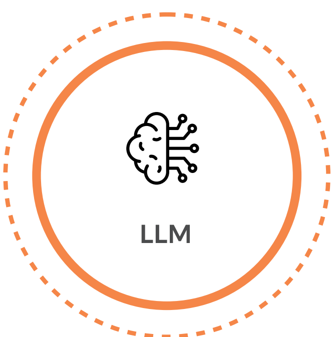
Goal two: high-risk case identification.

Cost prediction is one problem. Identifying which cases carry high uncertainty around that prediction is another, harder one. We are building a second layer that flags cases where the prediction interval around our predicted cost is unusually wide, signaling to underwriters that even though we have given them a number, they should treat it with caution and account for it in the risk reserve portion of their proposal. This is the harder part of the work, both because predicting variance is a second-order problem and because outlier detection on a small dataset is genuinely difficult. It is in active development.



Feature Importance

Feature	Value	Shapley
Self Care Needs	**5**	**0.25**
LOS	**13**	**0.13**
...	...	...
Gender	**M**	**-0.09**
Injury Type	**Burn**	**-0.15**



Interpretation / Output

Top factors that increased the prediction:

The model saw that this case had 5 self-care dependency needs. This pushed the estimate up the most among the SHAP factors shown ....

Goal three: explainable, plain-language outputs.

This is where the architecture really paid off. For every prediction the model produces, we extract Shapley values, which quantify how much each input feature pushed the predicted cost up or down for that specific patient. We then pass those values to a large language model whose only job is to translate the feature importance table into plain English.

The LLM is doing nothing clever. It is acting as a translator between the machine learning model's interpretability layer and the human readers who need to defend the prediction in a clinical conference or in contract conversations with the client. That is the point. The intelligence lives in the architecture, not in the model selection.

What This Looked Like in Practice

The full effort, from data access to three working capabilities, took roughly six weeks of substantive work. That includes data discovery on a fragmented, complex underlying dataset and contending with a development environment that was, frankly, hostile.

For context: the prior internal modernization effort lasted two years and was abandoned. We delivered three functional model components in six weeks because we did not start from "build a better model." We started from "map the workflow, understand the data constraints, and architect a solution that respects both."

> **We did not start from 'build a better model.' We started from 'map the workflow, understand the data constraints, and architect a solution that respects both.'**

The work is ongoing. We are still iterating on the high-risk identification piece, finalizing recommendations on how the outputs integrate into the existing intake process, and working through the change management question of whether the legacy one-through-six scale gets supplemented, replaced, or run in parallel during a transition period.

What We Took Away

A few lessons from this engagement worth carrying into any AI initiative in healthcare or insurance:

Small data is the rule, not the exception. In specialty insurance and complex care, you will rarely have the hundreds of thousands of training samples that consumer-facing AI use cases enjoy. Architectural choices, including what you predict, how you decompose the problem, and where you place the interpretability layer, matter more than algorithm selection.

Interpretability is not a feature. It is a precondition. In any regulated, high-stakes decision context, a prediction without an explanation is operationally dead. The architecture has to produce explanations as a first-class output, not as an afterthought.

> **A prediction without an explanation is operationally dead.**

The model is rarely the hard part. The hard parts are: getting clean data, fitting the AI into a human workflow that has rules, committees, and override mechanisms, and earning the trust of clinical and actuarial reviewers who have seen modernization efforts fail before. Architecture, informed by deep industry context, is what bridges those gaps.

This is what we mean when we describe X by 2 as an architecture-first AI consultancy. The AI matters. But the architecture is what makes the AI useful.

About the Author

Mitch Quinn
Director of AI/ML

Mitch Quinn is the Director of AI/ML at X by 2. With over 25 years of experience applying state-of-the-art research, Mitch is leveraging deep learning methods to solve the complex, critical problems facing healthcare. He has created and developed award-winning AI solutions to revolutionize engagement, equity and impact in care management and holds several patents in the fields of anomaly detection and classification.

X by 2 is a technology consultancy specializing in architecture-led transformation for the healthcare and insurance industries. We help leaders deliver AI, data, and platform initiatives that hold up under the operational, regulatory, and clinical realities of these industries. To discuss how this approach could apply to your organization, contact us.

Let's get to work.

Every successful partnership starts with an introduction. Send us a message or schedule time to chat.
contact@xby2.com