BESTOW

Right-sizing AI for life and annuity carriers: why the biggest model is rarely the right one

Published: September 2026

For the latest information, please see www.bestow.com/new-business

Table of Contents

The reflex that's draining AI budgets

For the last few years in the insurance space, the strategy around enterprise AI has been a pretty simple binary. Either employ the biggest, most powerful model available, or find the cheapest, "lowest-risk" model. When capability was scarce and usage was comparatively low, reaching for the most powerful model made sense. It was the safest bet, and when managed properly, the bill stayed small enough that nobody looked closely. On the other side of the coin, horror stories about runaway costs pushed some insurers toward low-risk, low-reward tools that actually didn't deliver very much of anything.

Both of those strategies are flawed in the same way. They start from the model instead of the job. It's a Goldilocks problem of sorts. One model is too cheap and not capable enough to trust. Another is too powerful and too expensive to justify. The right one for a given task sits somewhere in between. And finding the model that's "just right" is the whole game.

Today, capability is abundant across a wide range of model sizes, and usage has scaled to the point where the bill, and the ROI behind it, is very much worth a look. What once seemed to be a sensible default strategy, whether you aimed high or low, has quietly become a major driver of avoidable AI spend on the enterprise balance sheet.

You've probably heard of "tokenmaxxing," the leaderboard culture in which AI consumption becomes the scoreboard and burning more tokens gets treated as proof of productivity. Carriers face the enterprise version of the same conflation: mistaking model power and AI spend for AI value. Inside a life insurance business, that confusion doesn't show up as a novelty dashboard. It shows up in operating costs, as stalled or failed projects, and in a widening gap between what carriers spend on AI and defensible ROI.

The truth is that most of AI's returns in life insurance won't be decided by model power. They'll be decided by discipline: start from a defined business outcome, then choose the fit-for-purpose tool rather than defaulting to the cheapest or the most expensive one. The carriers that pick capability first and go hunting for a use case second are at real risk of spending the most and showing the least.

The bill nobody modeled for

Let's talk about inference. That's the cost of running a model in production, as opposed to training it once, and it has become one of the two largest line items in enterprise AI budgets, behind only talent. Unlike a software license, it doesn't sit flat on the books. Instead, it scales up with every query, every retrieved document, and every step an agent takes. The more the organization uses AI, the more it pays.

What catches finance teams off guard is that the unit price of inference is falling fast, roughly an order of magnitude over the past two years for a given capability tier. The intuitive conclusion is that AI is getting cheaper and the cost problem will solve itself. That's not entirely wrong. Some prices have dropped (though some have gone up), models have gotten more capable, and some of that will continue. But prices won't fall forever, the models won't get infinitely smarter, and we're already seeing a taper in that curve. At the same time, token-per-task is rising faster than price is falling. A simple 2024-era chatbot exchange might have consumed a couple of thousand tokens. A 2026 agentic workflow that plans, calls tools, checks its own work, and retries can burn tens of thousands of tokens on a single transaction. When the price per token drops tenfold but the tokens consumed per task rise twenty-five-fold, the invoice goes up, not down.[3] A multi-million-token month (or week, even) is the new normal for an AI-first developer. The cost to complete the same task tends to fall over time. The trouble is that the tasks AI gets asked to do keep growing in complexity, in token usage, and thus in cost, as well.

That's the case for discipline. Right-sizing isn't a way to squeeze out a little extra savings while AI gets cheaper on its own. It's the only lever that keeps the bill from climbing as usage deepens. A carrier that routes every task to a frontier model pays a premium on every routine decision that never needed one, and it pays that premium more times each quarter. There's a temptation to double down, because each new model offers better intelligence at the same price, which makes reaching for the best one seem like a bargain. But what used to require the top-of-the-line model may now run fine on a mid-tier model. Sometimes the same budget just buys you a harder task done well.

But cost is only the first pressure. The second is governance. You can see the tension in Grant Thornton's 2026 AI Impact Survey, which drew on roughly 100 respondents across the insurance sector within a broader panel of 950 executives. Insurers are getting real returns; about 52% reported AI-enabled revenue growth, fifteen points above the cross-industry average.[1] Yet the same group can't yet prove the systems behind those returns are under control. To that point, 44% said governance or compliance challenges had contributed to an AI project failing or underperforming, and only 24% were very confident they could pass an independent AI governance review within 90 days.[1] Real revenue on one side, unproven governance on the other. That's the ROI problem summed up.

Meanwhile, there's a third pressure that actually comes from inside the industry's own underwriting logic. Insurers have started pricing unbounded generative AI as a risk to be contained. In 2026, standard commercial general liability forms introduced dedicated generative-AI exclusion endorsements, and specialty insurers began attaching sublimits to AI-related exposure.[2] Sit with that irony for a second. The businesses whose entire discipline is pricing risk looked at open-ended generative AI and decided to cap their own exposure to it. Any carrier deploying that same class of technology inside its operations should take the hint and ask where "unbounded" is the wrong setting within its own technology stack.

The discipline: define the outcome, then right-size the tool

None of this is an argument for small models over large ones. That's just the same problem run in reverse, and it fails for the same reason: it starts from the tool. The best path forward is a one-way street. Define the business outcome first, then deploy the lowest-cost tool that clears the reliability, regulatory, and accuracy bars for that specific job, whether that turns out to be a distilled small model, a fine-tuned mid-tier one, or a complex frontier system.

The factors that determine what model will work best for a given task or process are evolving. You can now get intelligence approaching Claude Fable 5 or GPT-5.6 Sol from inexpensive-to-license models coming out of China, such as Moonshot's Kimi K3.[4] The open weights are attractive to a regulated carrier, since you can keep sensitive data in-house and run the kind of audits, red-teaming, and traceability that closed APIs won't allow. But “open” doesn't mean cheap to run. A model like K3 carries 2.8 trillion parameters and demands serious infrastructure to host securely, and it brings baggage a U.S. carrier can't wave away: geopolitical risk, the possibility of model bans, reputational backlash. So the "right" model has become a moving target, defined all at once by cost, capability, compliance, security, and geopolitics.

In practice, think of right-sizing as orchestration. Route each task to the model that fits it. A narrow predictive model where the job is a well-defined prediction. A compact generative model where the job is bounded language work. A frontier model held in reserve for the genuinely hard, ambiguous, high-stakes reasoning that earns its cost. The escalation path is deliberate and calibrated. On tasks where a smaller model can do the work in place of a frontier model, it can be about ten to thirty times cheaper to serve.[5]

There's a second reason routine tasks make good candidates for cheaper models. On work you were going to check anyway, you don't need to pay frontier-model prices for confidence you'll verify yourself. If a task already has a review step built in, whether that's a human sign-off, a downstream validation, or a reconciliation, then a lower-cost model doing the first pass is often the smart economic call. The check catches anything the cheaper model might have missed.

The same logic drives a pattern some teams call "10-80-10." Use the expensive model to plan (about 10%) and to review (about 10%), and let lower-cost models handle roughly 80% of the execution under the frontier model's direction. Done well, that can cut costs by half or more and still deliver results.

Push the idea one step further and it changes how carriers should think about generative AI altogether. The expensive, general-purpose model, and the 10-80-10 pattern, is often best used to build rather than to run. Spend frontier-model capability once to design and stand up a contained application, then run that application more cheaply and predictably every time the task recurs, and you stop paying frontier prices on every single execution. Generative AI stops being the engine and instead becomes the tool that builds the engine. Picture an internal analytics capability that notices when a key metric moves and can help surface reasoning behind that move, tracing a conversion-rate dip to a possible cause without anyone needing to know the right question to ask in advance. The heavy model does that work once, figuring out the signals and parameters worth watching. The capability it leaves behind runs much leaner.

For a CRO, all of this reduces to one AI measurement habit: track cost per task and per workflow, not cost per token. Cost per token is an infrastructure number, while cost per task is a business number. The latter tells you whether a workflow earns its keep, and it turns model selection from an engineering preference into an economic decision the business can actually justify and govern.

Proof: right-sizing in production

The principle is clear, but while it's easy to wax philosophical about, it's harder to live. Here are some examples of what this looks like operationally at Bestow.

A recommendation engine built on narrow AI.
Bestow's recommendation engine is a deliberate case of matching the tool to the task. It runs on narrow AI, a purpose-built predictive data model rather than a general-purpose generative one, and it earns its place on economics, not novelty. Served to just 20% of traffic for one carrier, it drove roughly a 22% reduction in per-policy underwriting costs, a comparable reduction in cost per submitted application, and a 5% lift in approval rate, hitting up to 83% approval-prediction accuracy from only five customer-provided data points. Concept to implementation took about a month. The reason it works is because a contained, task-specific model is cheaper to run than a frontier model on the same job, and it's easier to govern and to explain to a regulator. Given the pressures above, that second part is no small thing.

An IVR system that contains AI where containment is the point.
Bestow's IVR runs in production on live customer calls. It authenticates callers and pre-populates agent screens before the conversation even starts, saving an estimated 30-60 seconds per call in handling time across tens of thousands of monthly interactions, with a roadmap toward contained self-service actions like adding a beneficiary or updating a name. This is right-sizing in a customer-facing setting. The system is scoped to do specific, verifiable things well, instead of turning an open-ended model loose on sensitive policyholder interactions. That bounded design is what makes the automation trustworthy enough to put in front of a real caller.

Two different classes of AI, each chosen the same way: from the job backward.

Why right-sizing points to a platform

Here's the uncomfortable part of everything you've just read. Right-sizing isn't a decision you make once. It's a job that never ends. The AI goalposts move every few weeks. A model that was the right call last quarter gets undercut by a cheaper one this quarter, or surpassed on capability, or flagged for a compliance or geopolitical reason nobody saw coming when you picked it. Doing this well means watching the market continuously, re-benchmarking, re-checking the cost-versus-risk math, and rewiring workflows when the answer changes. For a carrier whose primary focus is managing a large book of business, that's probably not the most strategic area to direct resources toward.

So how do carriers navigate right-sizing on their own? The reality is, they don't have to — at least not for every use case. That's the case for using AI as part of a platform you already rely on for things like underwriting, policy admin, billing, claims, etc. The strongest reason to run AI as part of a platform is that a good platform solution is already doing the model-fit work for you. In fact, they're doing it continuously, as one of their core jobs. They watch the market so you don't have to. They swap in the right model for each task as the options shift. And this isn't about surrendering control. The best platform of the future does the opposite. It gives you a menu of models and the expertise to navigate it, so you can pick the one that fits a given task rather than being locked into a single house model for everything. Because there is no one model to rule them all, and there likely won't ever be. There's the choice of which model, and then, within a model, which version best fits the job while keeping spend down and margin up. An effective platform partner brings the data to guide that call and keeps guiding it as the options multiply and mature. And in a SaaS arrangement, the vendor is commonly the one paying for the inference. They're motivated to use the lowest-cost model that still delivers what customers need. Not the priciest option, and not some cut-rate one that fails and floods support with tickets.

They win when they right-size, and you get the benefit without having to open the toolbox.

We've seen this movie before. Not long ago, every serious company assumed it had to run its own infrastructure, and that owning their servers was the only way to get things just right. Enter: the cloud. It turned out that letting someone else manage all that was easier, more reliable, and, when done right, no more expensive. Today almost nobody questions that trade. AI is heading the same direction. Right-sizing across a fast-moving model landscape is exactly the kind of specialized, evolving work a platform partner should own for many use cases, the same way you'd never build your own data center just to run company email. So while you could, for example, build your own intelligent document processing solution for underwriting or claims documents, that doesn't mean you should.

The distance from demo to production

Another way the ROI math can go sideways is when a prototype comes easy, but enterprise-grade software does not. That's the gap between something that works as a demo and something that works at 2 a.m. on a holiday weekend, unattended, for the one millionth time.

The industry has learned this one the expensive way. Getting to a convincing prototype has never been easier or cheaper. A small team can stand up something impressive-looking in a matter of days. But a prototype is not a product. The distance between a demo that dazzles in a conference room and enterprise-grade software that runs day in, day out, through edge cases and outages and audits and regulatory change, is enormous. Much of the AI spend in our industry right now is buying the first thing while assuming it will become the second. It usually doesn't, at least not on the timeline or budget anyone planned for.

This is where the "R" in ROI tends to quietly disappear. You can spend a great deal of money getting to a demo and still have nothing generating a return, because a demo doesn't underwrite a policy or pay a claim or answer a call at scale. The return only shows up once you have the durable, production-grade version, and the journey to that version is longer and harder than the prototype makes it look.

This isn't an argument that carriers should never build anything themselves. Plenty of AI work is well worth doing in-house: automating a back-office task, standing up an internal dashboard, spinning up a chatbot for an HR team. Those are real and useful, and no one needs a platform provider's permission to pursue them. The distinction that matters is core versus peripheral. Building your own agentic tools to smooth over point-solution gaps or to make software do what it should have done in the first place is a different thing entirely. That's papering over the problem instead of addressing the underlying cause. The core systems that run the business are where the demo-to-production gap is most punishing, and they're precisely the place to lean on a platform rather than a prototype.

None of this is a reason to be timid about AI. The returns are real, and when the work is done right, they're substantial. That's precisely the point: there is a very healthy return to be made here, but it's made by getting all the way to production and sustaining it, not by stopping at the demo. That last mile, done well, is a large part of the business Bestow is in.

Conclusion: what disciplined carriers do now

The path forward isn't complicated, but it does require a bit of unlearning.

For three years the reflex has been to start with the model and find a use for it. The discipline is to reverse that: start from the outcome you're after, and reach for the least expensive tool that passes muster. Taking an outcome-first, model-second approach fixes most of what's broken about how many carriers are currently spending on AI.

The rest follows from being honest about what's worth owning. Right-sizing models as the market churns, and carrying a prototype the long distance to production-grade software, are full-time jobs. Neither one underwrites a policy, pays a claim, or grows a book of business.

The carriers that win with AI over the next decade won't be the ones that spent the most, or ran the largest models, or stood up the biggest in-house AI teams. They'll be the ones that picked the right platform partner to make these calls well, again and again, and spent their own energy where it actually moves the business.

[1] Grant Thornton, 2026 AI Impact Survey Report

[2] ISO 2026 commercial general liability generative-AI exclusion endorsements.

[3] "Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025" (Gartner, 2026).

[4] 2026 Kimi K3 Benchmarks, Capabilities & Access Guide

[5] "Small Language Models are the Future of Agentic AI" (NVIDIA, 2025)

BESTOW

Bestow Inc.
enterprise@bestow.com

2700 Commerce St, Ste. 1000
Dallas, TX 75226