braven

INTRODUCTORY GUIDE · UPDATED AUGUST 2026 · 25 MINUTE READ · NO TECHNICAL BACKGROUND ASSUMED

An introduction to AI for insurance operations

Written by Carlos Chávez,
Co-founder and CEO of Braven.

braven

An introduction to AI for insurance operations

Written by Carlos Chávez, Co-founder and CEO of Braven.

Braven is headquartered in San Francisco with an office in London, and works with delegated authority teams across seven countries.

Published by Braven, the AI operations platform for MGAs and reinsurers that do their own underwriting, at gobraven.com. Reviewed quarterly.

Introductory guide · Updated August 2026 · 25 minute read

Contents

Sources, page 44.

01

It started with ChatGPT

Almost every conversation I have about AI with an MGA begins in the same place. Somebody on the team was using ChatGPT.

Almost every conversation I have about AI with an MGA begins in the same place. Somebody on the team was using ChatGPT. It started with proof reading and later became an informal part of their workflow. They were impressed, or unnerved, or both. And they cannot work out why the software they were shown last week, which cost considerably more, appeared to do so much less.

What you have actually been using

Three products dominate. ChatGPT is made by OpenAI, Claude by Anthropic, Gemini by Google. Behind each sits a large language model (LLM), and behind that sits a training process that consumed a very large proportion of the written internet.

Almost every AI product being sold into insurance, including Braven, is built on top of one or more of these rather than on a model the vendor trained themselves. Training a frontier model costs hundreds of millions of dollars. Nobody in insurance technology is doing it and nobody in insurance would want to. So, when a vendor says “our proprietary AI”, what they almost always mean is “our software wrapped around somebody else’s model”. This is a sensible way to build. It just means the interesting question, the value they uniquely deliver, is what they built around it.

What “large language model” means

Large describes the size of the thing. What a model learned during training is stored in numbers called parameters, and the current frontier models have hundreds of billions of them. Nobody, including the people who built it, can point at a particular parameter and say what that one does.

Language is both what it was trained on and what it produces. In “AI speak”, program code is as much of a language as a form is. So is a schedule of values, which is why this generation of software can read one and the last generation could not.

Model means a statistical system rather than software somebody wrote line by line. Nobody coded a rule saying that a slip usually contains an attachment point. It picked that up from reading a very large number of slips.

What happens when you ask an AI chatbot something

You type a question. Back comes an answer: fluent, well organized, delivered without hesitation. Where did it come from? Well, a language model has been trained to do one thing: given some text, work out what text should come next. That's it.

Do it at sufficient scale and something unexpected emerges, because to predict the next word in a sentence about excess of loss reinsurance, a system has to build up an internal representation of how excess of loss reinsurance works. Prediction, pushed hard enough, starts to resemble understanding. But it remains a prediction. Nothing was looked up. No database consulted, no source checked. What you received was the most probable continuation of your question, given everything the model had absorbed during training.

This is why it never says "I don't know."

Tokens, compute, and where your data actually goes

All that prediction comes at a cost. And, in the world of AI, the currency of the day is the “token”.

Tokens. A model doesn’t read words, it reads tokens: fragments of roughly three or four characters. A hundred page schedule of values might come to forty thousand of them. Nearly all AI pricing is per token, so a heavy renewal season costs more than a quiet month. And every model has a ceiling on how many tokens it can hold at once, called the context window, which is why a very large document sometimes has to be handled in pieces rather than all at once.

Compute. Running a model is not free. Every question consumes time on specialized chips, and somebody is paying for the electricity. This is worth remembering when a vendor offers unlimited usage: unlimited use of something with a real unit cost usually means a limit exists somewhere and you have not been shown it. Ask what happens to the bill when your submission volume doubles.

Data centers. All of this processing happens on somebody else’s computers, in a physical building, in a particular country. When an underwriter pastes a slip into a chat window, that slip leaves your office. For most businesses that is a mild curiosity. For a delegated authority business it is the question your capacity provider will eventually ask, and the one your data processing agreements already cover. “Where is our data hosted, and can we choose the region?” has a physical answer.

Prediction, pushed hard enough, starts to resemble understanding. But it remains a prediction.

02

Why AI sounds certain even when it is wrong

That inability to say “I don’t know” is the most consequential thing in this guide.

That inability to say “I don’t know” is the most consequential thing in this guide.

> **Try this experiment yourself:** Open ChatGPT and ask it about the attachment point on one of your own treaties. Name the cedent and be specific.

It will answer. The answer will be well formatted, plausible, and completely invented, because the model has never seen your treaty and has no mechanism for telling you so. The model is doing exactly what it does: producing text that looks like the right kind of answer.

This is what people mean by hallucination. To me, it’s a poor word, because it implies malfunction, and this is not one. AI built to produce plausible text will produce plausible text, and plausibility and accuracy are simply different properties. Nobody is going to fix it in the next release.

People reach for the same comparison: a very well read graduate on their first day. Useful, as far as it goes. They have read more than anyone in your office. They write quickly and clearly. They will work through the night without complaint.

But an actual graduate, asked about a treaty they have never seen, says so. This one will not, because composure is the thing it was trained to produce. For a regulated industry like insurance, this introduces meaningful risk.

Composure is the thing it was trained to produce.

The fix is not a better model

The answer to hallucination is not waiting for the next version. It is grounding: giving the model the actual document and requiring it to answer from that rather than from training.

Ask the same question with the treaty attached and the behavior changes completely. Now it is reading rather than remembering.

Grounding is the single most important safeguard in any document heavy business, and it has a visible signature you can check for: in a properly grounded system, every extracted value links back to the page it came from. You should be able to click the attachment point and land on the slip, at the right line. It takes seconds to verify. An ungrounded number has to be checked from scratch, which means the software saved you nothing at all.

03

What is an AI agent?

No word in this market is more abused.

No word in this market is more abused.

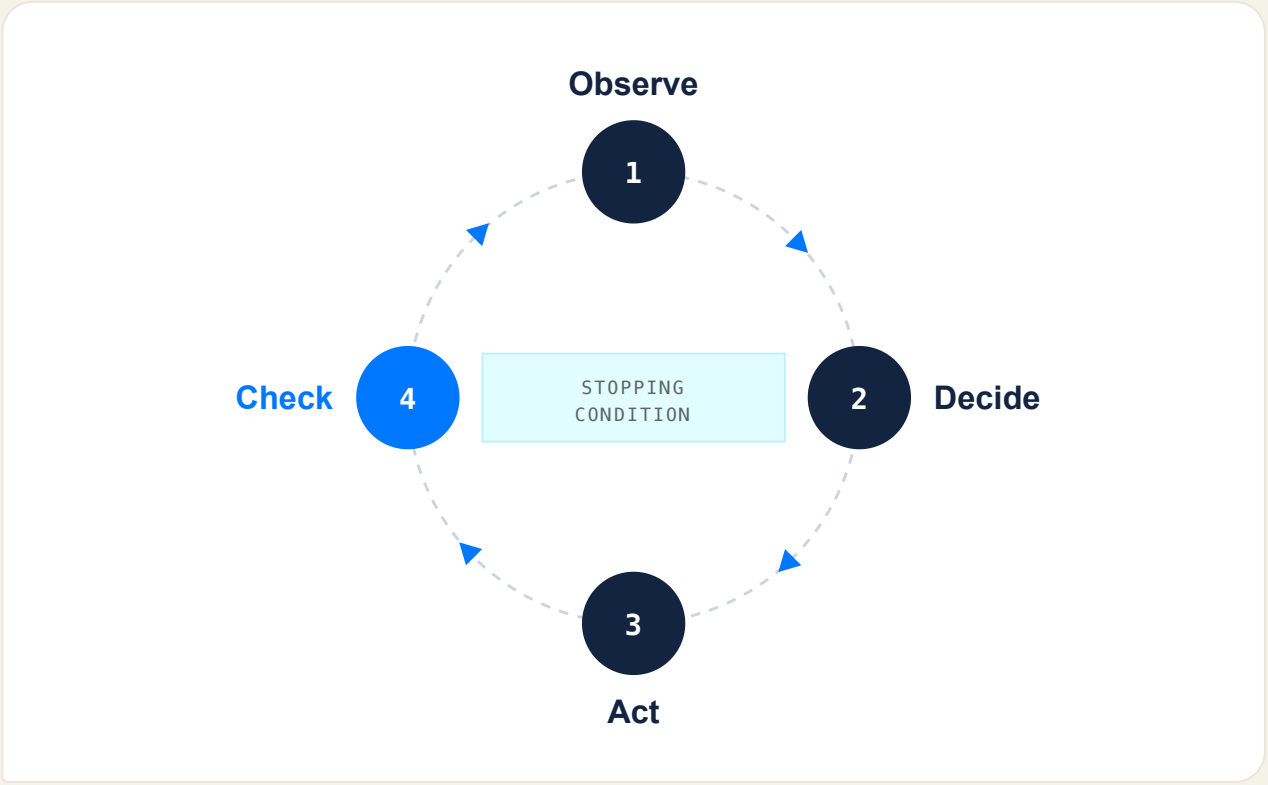
Your ChatGPT session was a single exchange. You asked, it answered, it stopped. Nothing in that arrangement lets it check whether the answer was any good, or do anything about it if not.

An agent is a model that can run in a loop and use tools. Practitioners use a shorthand: an agent equals a model, plus memory, plus planning, plus tool use.[5] Those four together let it work toward a goal across several steps rather than answering one question and going quiet.

Tool use, sometimes called function calling, is a model's ability to do things beyond producing text: Query a database. Send an email. Run a calculation. It is the bridge between a system that talks and a system that works, and it is also precisely where permissions stop being theoretical, because a tool that can send is a tool that can send the wrong thing.

The loop is the cycle it runs while working:

FIGURE 1 · THE AGENT LOOP



1. **Observe.** Take in what is there: a new email, the result of the last action, an error.
2. **Decide.** Work out what to do next, given the goal.
3. **Act.** Do it. Read the document, query the system, draft the reply.
4. **Check.** Did that work? Is the job finished? If not, go round again.

“Summarize this submission” is a single exchange. “Watch this inbox, and when a submission arrives, read it, check it against appetite, and either decline it with a reason or draft a quote and flag it for review” is an agent, because it involves several steps, decisions that depend on what it finds, and a condition for stopping.

Reasoning vs. Work

An agent has two layers: the model, which reasons, and the harness, which is the surrounding software that assembles context, calls the tools, enforces the limits and keeps track of where it is.

THE MODEL

Reasons

THE HARNESS

Assembles context, calls tools, enforces limits, keeps state

Practically all the engineering, and nearly all the difference between one product and another, lives in the harness. Two vendors running the identical model can produce systems that behave nothing alike. So, when somebody tells you which model is underneath, they have told you very little about what they built.

5 degrees of autonomy

"Agent" covers a range rather than a setting. Here's how I see it applied across insurance:

DEGREE	WHAT HAPPENS	AN UNDERWRITING EXAMPLE
0. You do it	No AI involved	Reading the submission yourself
1. It suggests	The system proposes, you decide everything	A draft decline letter you rewrite
2. It drafts, you approve	Work is finished but nothing takes effect without a person	A quote prepared, checked against binder, waiting on your click
3. It acts, you supervise	The system acts inside set limits; a person watches and can intervene	Chasing a missing loss run and logging the replies
4. It acts, you audit	The system acts, a person reviews afterwards on a sample	Triaging out submissions clearly outside appetite

No degree is right in general. However, as a rule, autonomy should fall as consequence rises. Chasing a follow up sits at 3 or 4. Binding a risk should never be above a 2, and a vendor who suggests otherwise is telling you how little delegated authority work they have actually seen.

Why most of this is oversold

Gartner reckons that of the thousands of vendors claiming agentic AI, only around 130 genuinely have it, and coined the term “agent washing” for the rest.[3] It also expects more than 40% of agentic AI projects to be canceled by the end of 2027, chiefly because of cost, unclear value and inadequate controls.[3]

130

of thousands of vendors claiming agentic AI genuinely have it[3]

40%

of agentic AI projects expected to be canceled by the end of 2027[3]

You do not need to referee that. One question does most of the work:

What does it do when nobody is asking it anything?

If the answer is “it waits”, you are looking at an assistant. Possibly a good one, possibly worth buying. Not an agent.

04

The four kinds of AI in insurance, and where each belongs

Marketing lumps these together. They are different technologies that fail in different ways, and knowing which is which will save you money.

Marketing lumps these together. They are different technologies that fail in different ways, and knowing which is which will save you money.

1 Rules engines, which are not AI at all

If this, then that. Rules that are written by a person and executed identically every time. Your rating engine is probably one, and so is most of what gets sold as workflow automation.

The distinction is legal as well as technical. Under the EU AI Act, an AI system is one that infers from its inputs how to generate outputs, with some degree of autonomy.[1] Hard coded rule based software generally falls outside that definition entirely.[2]

WHERE IT FITS

Anywhere you need the identical answer every time. Sanctions screening. Limit checks. Referral triggers. Do not let anyone replace these with a language model. You would be trading determinism for fluency, which is a poor bargain in compliance.

2 Predictive machine learning (what your actuaries have used for years)

Models trained on historical data to predict a number or a category. Loss cost, propensity to lapse, likelihood of fraud. As the header suggests, the concept is neither new nor controversial: close to 80% of insurers in WTW's 2026 survey use advanced rating and pricing models, with another 11% planning to.[6]

Close to 80%

of insurers surveyed use advanced rating and pricing models[6]

11%

more are planning to[6]

WHERE IT FITS

Anywhere you have structured historical data and want a number out of it. Keep your pricing model as a pricing model. Language models are the wrong instrument for that job and always will be.

3 Generative AI

It produces text, structured data, code and images. In insurance, the application that matters is reading and writing documents.

Adoption was fast: More than half the property and casualty insurers in WTW's 2026 survey already use large language models or generative AI, with a further 29% planning to inside two years.[6] Sollers puts four in ten insurers using AI somewhere in underwriting, and roughly 20% of commercial insurers already using it to triage submissions and pull data out of unstructured documents.[8]

WHERE IT FITS

Unstructured language. Every previous generation of software could only handle the tidy ~10% of your inbound and left the rest to people. This is the first technology that can read the other 90%.

4 Agentic systems: the current frontier

Generative AI plus the loop and the tools from section 3. Not a cleverer model, a different architecture around one.

WHERE IT FITS

Multi-step operational work with a clear finish line. Getting a submission from arrival to a decision, with the evidence attached.

Putting each where it belongs

A serious MGA operation can run all four: rules for the things that must never vary, predictive models for the numbers, generative AI for the documents, and agents for the sequences.

Be wary of any vendor proposing to replace a category that already works. Rules engines being swapped for language models is the most common expensive mistake in this market, and it tends to surface during an audit.

FIGURE 2

Where each kind belongs

They are different technologies that fail in different ways.

NOT AI **Rules engines**	Anywhere you need the identical answer every time. Sanctions screening. Limit checks. Referral triggers.
DECADES OLD **Predictive machine learning**	Anywhere you have structured historical data and want a number out of it.
THE RECENT ARRIVAL **Generative AI**	Unstructured language. Every previous generation of software could only handle the tidy ~10% of your inbound and left the rest to people. This is the first technology that can read the other 90%.
THE CURRENT FRONTIER **Agentic systems**	Multi-step operational work with a clear finish line. Getting a submission from arrival to a decision, with the evidence attached.

A serious MGA operation can run all four.

05

Where AI excels... and what it does poorly

Confidence is not calibrated to correctness.

Genuinely good at

Reading things never designed to be read by software.

Scanned PDFs, phone photographs, spreadsheets with merged cells and a hidden tab, a slip in a format that quietly changed at last renewal. This is the capability that brought the technology into insurance rather than passing it by.

Turning language into fields.

Occupancy, construction, total insured value, attachment point, cedent, inception date. Pulled out and put where they belong. The industry calls this extraction, and it is the workhorse.

Sorting and routing.

This one is in appetite, this one is not, this one needs the marine team, this one is a renewal of something you declined last year.

Comparing one document against another.

Submission against appetite. Bordereau against binder. This year's wording against last year's. Tedious work, error prone when a tired person does it at five o'clock, well suited to a machine.

Drafting.

Quote letters, decline letters, broker follow ups, the covering note. Not final copy. A strong first pass.

Working in the gaps between systems.

Most operational time vanishes in the space between your inbox, your policy administration system and your spreadsheets. That space is where this technology is most useful, and it is also the space no previous software could occupy.

Genuinely bad at

Arithmetic.

Counterintuitive, but a language model predicts what a calculation looks like rather than performing one. Well built systems hand the numbers to a calculator instead of doing them in the model. Ask any vendor how they handle it; the answer is a good tell.

Knowing when to stop.

Covered in section 2 and worth repeating, because it is the failure that costs money. Confidence is not calibrated to correctness.

Judgment under real uncertainty.

Whether to back a broker's account of a loss that looks worse than it is. Whether a class is turning. Whether to write something marginal to hold a relationship. These are the reasons your underwriters are worth what you pay them, and nothing here changes that.

Being accountable.

Models cannot hold authority and nobody can delegate responsibility to one. Your capacity provider will hold you responsible for every decision made on their paper, whatever produced it.

Consistency, unless constrained.

Same question, different answers. Fine for prose. Not fine for anything a regulator might sample.

Your appetite.

Unless you tell it. It has no idea what you will not write, and will cheerfully assume the market's general appetite is yours.

Models cannot hold authority and nobody can delegate responsibility to one.

06

Permissions, evidence, and who is accountable

With AI, we're essentially talking about the same conversation applied to a different kind of employee.

Nobody joins your firm with full binding authority. They get scoped permissions, their work gets checked, and there is a record of what they did. With AI, we're essentially talking about the same conversation applied to a different kind of employee.

Four things do most of the work.

1 Permissions inherited from a person

Every automated action should run on behalf of a named human and never exceed that human's authority. If an underwriter cannot bind above $5 million, nothing acting for them should either.

Obvious, and regularly not the case, because it is easier to build a system with broad service level access than to plumb it through your permission model. Consider asking: can an automated action ever exceed the permissions of the person who triggered it? Any hesitation is telling.

2 An audit trail you did not have to assemble

Every action being recorded, human and machine alike. Every extracted field linked to its source. The complete trail for one risk retrievable on demand rather than reconstructed from three inboxes and somebody's laptop.

MGAs consistently undervalue this and capacity providers consistently ask about it first. Get it right and your next binder review becomes a query rather than an event.

3 The right oversight for each action

Go through your workflow and decide, action by action, what needs approval before it happens, what needs supervision while it happens, and what can be sampled afterwards. Write it down. This simple document represents most of what will become your AI governance policy and preparing it will take just a few days.

BEFORE	DURING	AFTER
Approval before it happens	Supervision while it happens	Sampled afterwards

4 Data handling you can point to in a contract

The answers to 4 simple questions: (i) Is our data used to train anyone's model? (ii) Where is it hosted? (iii) What do we get back if we leave? (iv) Which sub processors touch it, and how are we told when that list changes?

Consumer tools generally answer these badly for business use. Having met hundreds of MGAs, I can say this is not an argument for banning them, because bans push the behavior somewhere you cannot see it. It is an argument for a one page acceptable use rule, and then for making the sanctioned system good enough that nobody needs to reach around it.

The frameworks worth knowing

Finally, it's worth being familiar with the two frameworks you will hear referenced:

NIST's AI Risk Management Framework is the American reference point. It's voluntary, organized around four functions: govern, map, measure, manage.[4]

ISO/IEC 42001 is the certifiable equivalent, for when a counterparty wants an audited answer rather than your word.

Neither is required reading for an MGA. Both are useful vocabulary when a carrier's compliance team asks how you manage this, and being able to name them often changes the temperature of that conversation.

07

Where to begin

You can kick things off at your MGA without a budget.
Here's where I would start:

You can kick things off at your MGA without a budget. Here's where I would start:

1 **Find out what is already happening.** Survey your team, with no consequences attached, what they already use. The answer is rarely nothing. Insurers consistently report their obstacles as data readiness, security and legacy integration rather than the technology itself,[7] and unmanaged use is where those risks are sitting right now.

2 **Write a one page acceptable use rule.** No client data, no submission documents, no loss runs, no personal data into any tool your data processing agreements do not cover.

3 **Measure your operation before you change it.** Time to quote. Submission capture rate. Hit rate. Bordereaux production time. Without a baseline you will always be arguing about feelings.

4 **Pick just one flow (not a department).** The projects that work start with just one flow.[9] Submission intake is usually the best first choice, being high volume, unstructured, and currently consuming your best people's mornings.

5 **Then go shopping with the criteria you wrote yourself.** Our companion piece, *The MGA's buyer's guide to AI operations*, has a weighted scorecard, forty questions to send vendors, and a ninety day evaluation that ends in a decision.

And, if you want to merely see an AI Operations Platform working on your own documents, send us your messiest submission and we will run it in thirty minutes at no cost, at gobraven.com.

Glossary

Every entry includes what the term means and what it looks like in a delegated authority business.

Agent Software that pursues a goal across multiple steps, deciding what to do next and using tools along the way, rather than answering a single question. Commonly described as a model plus memory, planning and tool use.[5] The test: what does it do when nobody is asking it anything?

Agent washing Marketing an assistant, chatbot or scripted automation as an autonomous agent. Gartner estimated only around 130 of the thousands of vendors claiming agentic AI were the real thing.[3]

Agentic AI AI built to act toward goals across several steps rather than respond to single prompts. The label alone tells you nothing about capability.

AI system Under the EU AI Act, a machine based system operating with some degree of autonomy that infers from its inputs how to generate outputs such as predictions, content, recommendations or decisions.[1] Rule based software generally falls outside this definition.[2]

Audit trail The record of every action taken in a system, human and automated, retained so a single decision can be reconstructed later. The practical test: how long does it take to answer "why did we write this risk?"

Compute The processing power consumed when a model runs, and the reason nothing here is free. Relevant whenever a vendor offers unlimited usage of something that has a real unit cost.

Confidence score A system's own estimate of how reliable a particular output is. Useful only when it drives behavior, meaning low confidence produces a flag rather than a guess.

Context window How much text a model can hold in mind at once. Large now, but finite, and exceeding it drops earlier material, usually silently. A 400 page schedule plus a binder plus your appetite statement may not fit at once; how a system handles that has real consequences for accuracy.

Data residency Where your data physically sits and is processed. A contract question with a physical answer: which country, which provider, and whether you get to choose.

Extraction Turning unstructured content into structured fields. A scanned ACORD form becomes named fields in your policy administration system.

Extraction completeness How much of a submission file was captured, structured and made searchable. Different from accuracy, which is about whether what was captured is correct. Ask about both.

Fine tuning Additional training on specific examples to change how a model behaves. Changes style and format; does not teach it facts about your book. Fine tuning could teach your house style for quote letters. It will not teach it what is in this month's submissions. For facts, use grounding.

Frontier model The largest and most capable models available at any given moment, currently from a handful of providers. Training one costs hundreds of millions of dollars, which is why nobody in insurance is building one.

Generative AI AI that produces new content, including text, structured data, code and images, rather than only classifying or scoring what already exists.

GPU The specialized chip most AI processing runs on. You will never buy one. It explains why every question has a cost rather than being free once the software is installed.

Grounding Tying output to a specific source document so the model works from your material rather than its training. In practice: every extracted field links back to the page it came from. Grounded output is checkable in seconds; ungrounded output has to be checked from scratch.

Guardrails Constraints on what a system may do or say, whether technical, policy or contractual. Never bind without approval. Never quote outside appetite without a referral.

Hallucination Fluent, confident, false output. A property of prediction rather than a defect. A summary giving an attachment point of $5 million when the slip says $2 million, formatted so correctly that nobody checks.

Harness The software around a model that assembles context, calls tools, enforces limits and maintains state. Most of the difference between two products lives here rather than in the model.

Human in the loop A person approves before an action takes effect. Appropriate for anything that binds, prices or commits the firm.

Human on the loop A person supervises and can intervene, but the action does not wait for approval. Appropriate for chasing a missing loss run. KPMG treats escalation between the two as non negotiable.

Inference A single run of a model, taking input and producing output. Usually the unit you are billed for.

Large language model (LLM) A model trained on very large quantities of text to predict what text comes next. The technology underneath ChatGPT, Claude, Gemini and most current AI products.

Least privilege The principle that any actor, human or automated, holds only the permissions its task requires. In practice it means an automated action never exceeds the authority of the person it runs for.

Machine learning Systems that learn patterns from data rather than following written rules. Predictive models used in pricing and reserving are machine learning, and predate generative AI by decades.

Model The trained system that does the reasoning. Most insurance AI products are software built around a third party model. When a vendor says "our AI", ask whose model is underneath and what happens if that provider has an outage.

Multimodal A model that handles more than text, typically images too. Relevant because so much of your inbound arrives as photographs and scans.

Parameters The numbers inside a model that hold what it learned during training. Frontier models have hundreds of billions of them. The "large" in large language model.

Prompt The instruction given to a model. In a purchased product, usually written by the vendor and invisible to you.

Retrieval augmented generation (RAG) Searching a document set for relevant passages and handing them to the model with the question, so the answer comes from your documents rather than training data. "What does our binder with this carrier say about vacant property?" With RAG the system finds the clause and answers from it. Without it, the model invents a reasonable sounding clause.

Structured output Output produced in a fixed format, such as named fields, rather than free prose. What allows AI output to flow into a policy administration system.

Token The unit of text a model processes, roughly a word fragment. The basis of most usage pricing. If you are billed per token, a heavy schedule season costs more than a light one.

Tool use, or function calling A model's ability to act beyond producing text: query a database, send an email, run a calculation. The difference between a system that drafts a quote for you to copy out and one that puts the quote in your system.

Training data The material a model learned from. Distinct from the data you send it in use, and worth keeping distinct in any contract you sign.

Grounded output is checkable in seconds;
ungrounded output has to be checked from scratch.

Frequently asked questions

What is AI for insurance operations?

AI for insurance operations means using artificial intelligence to run the administrative and document work around underwriting: reading submissions, extracting data, checking risks against appetite and binders, drafting responses and building bordereaux. It is distinct from AI used for pricing or risk selection, which is usually predictive machine learning rather than generative AI.

What is an AI agent, in simple terms?

An AI agent is software that works toward a goal over several steps, deciding what to do next and using tools to do it, rather than answering one question at a time. The usual formula is a model plus memory, planning and tool use.[5] The practical test is what it does when nobody is asking it anything: if it waits, it is an assistant rather than an agent.

Is ChatGPT good enough for an insurance business?

For individual tasks, often yes, and plenty of capable underwriters use it that way. The limits are structural rather than about how clever the model is. It does not watch your inbox, hold your appetite rules, chase your follow ups or leave an audit trail, and it starts every session knowing nothing about your book. Separately, read the data processing terms before any client document goes near a consumer account.

What is a token, and why does AI pricing use them?

A token is the unit of text a model processes, roughly three or four characters. A word like “underwriting” is two or three tokens; a hundred page schedule of values might be forty thousand. Almost all AI pricing is charged per token, which is why a heavy renewal season costs more than a quiet month. Every model also has a limit on how many tokens it can hold at once, called the context window, which is why very large documents sometimes have to be processed in pieces.

Where is our data actually processed when we use AI?

On somebody else's computers, in a data center, in a particular country. When an underwriter pastes a slip into a consumer chat window, that document leaves your office and may leave your jurisdiction. For a delegated authority business this is a contract question rather than a technical one, and the answer belongs in writing: which region, which sub processors, and whether you are able to choose.

What is the difference between generative AI and machine learning in insurance?

Machine learning predicts numbers or categories from structured historical data, and insurers have used it in pricing and reserving for decades. Generative AI produces new content and is good at unstructured language: emails, slips, wordings, loss run narratives. Different tools for different jobs, and a well run operation uses both.

What is a hallucination in AI?

A hallucination is output that is fluent, confident and false, and it happens because a language model predicts plausible text rather than retrieving facts. In insurance the risk is specific: a summary stating an attachment point the slip does not contain, formatted so correctly that nobody checks. Grounding every output to a source document is the main defense.

Is AI accurate enough to read insurance submissions?

For most document types yes, but the better question is whether accuracy is measured and whether output is traceable. A field you can trace to a page is verifiable in seconds. A field you cannot trace needs the same manual check whether the vendor claims 90% or 99%. Ask how accuracy is measured, on what, and how often it is refreshed.

Can AI make underwriting decisions?

It can prepare them and it should not make them. Anything that binds, prices or commits the firm should require human approval, because accountability cannot be delegated to software. Your capacity provider will hold you responsible for every decision made on their paper regardless of what produced it.

Do we need to replace our policy administration system to use AI?

No. The work consuming underwriter time sits between systems, in email and documents, rather than inside the policy administration system. Any proposal starting with replacement has moved the risk onto you.

Is a rules engine artificial intelligence?

Generally not. Under the EU AI Act's definition an AI system infers from its inputs how to generate outputs and operates with some autonomy;[1] hard coded rule based systems typically fall outside that.[2] This matters practically as well as legally, because rules engines remain the right tool wherever you need an identical answer every time, such as sanctions screening or limit checks.

What are the risks of using AI in an MGA?

Four in practice. Hallucination, addressed by grounding output to source documents. Excess permissions, addressed by making automated actions inherit a person's authority. Absent evidence, addressed by an audit trail produced as a by-product of the work. And data leakage, addressed by contract terms on training, hosting and exit rather than by verbal reassurance.

How much AI knowledge does an underwriter need?

Enough to explain what your systems do and where they stop. That is roughly the standard the EU AI Act's literacy obligations already set for deployers, and a reasonable bar regardless of jurisdiction. The concepts in this guide are sufficient; nobody needs the mathematics.

Where should an MGA start with AI?

Find out what your team already uses, write a one page acceptable use rule, measure your current operation before changing anything, and pick a single flow rather than a department. Submission intake is usually the right first choice because it is high volume, unstructured, and currently consuming your best people's mornings.

How is AI regulated in insurance?

In the United States through the NAIC Model Bulletin, adopted by 25 jurisdictions including the District of Columbia, with California, Colorado, New York and Texas running their own frameworks. In the European Union through the AI Act, which treats life and health risk assessment and pricing as high risk. Property and casualty falls outside that particular provision and needs its own analysis. Our buyer's guide covers the current dates in detail.

Sources

1. Regulation (EU) 2024/1689 (the AI Act), Article 3(1), definition of an AI system.
2. European Commission, *Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689*, published 6 February 2025 under Article 96(1)(f) of the AI Act. Rule based systems, classical heuristic systems and basic statistical estimators generally fall outside the definition.
3. Gartner, *Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027*, 25 June 2025.
4. National Institute of Standards and Technology, *AI Risk Management Framework (AI RMF 1.0)*, January 2023, and the accompanying AI RMF Playbook.
5. Lilian Weng, *LLM Powered Autonomous Agents*, Lil'Log, 23 June 2023. The formulation "agent = LLM + memory + planning skills + tool use" is Weng's, and has since become standard in agent engineering practice.
6. WTW, *2026 Advanced Analytics and AI Survey*, March 2026. Based on 59 property and casualty insurers in the United States and Canada.
7. Insurance Journal, *Insurers cautiously navigate the next steps in AI adoption*, May 2026. Top reported challenges were data readiness (45%), security and privacy (43%) and legacy system integration (41%).
8. Jakub Śliwiński, Sollers Consulting, *Underwriting transformation: there is no silver bullet – only good decisions*, June 2026, sollers.com/en/insights/success-factors-in-digital-underwriting. Based on a Sollers survey across ten markets.
9. Aditya Challapally, Chris Pease, Ramesh Raskar and Pradyumna Chari, *The GenAI Divide: State of AI in Business 2025*, MIT Project NANDA, July 2025. Preliminary findings; the authors describe the figures as directionally accurate rather than audited.

braven

Architects of the Possible.
Move. Evolve. Win.

Published by Braven, the AI operations platform for MGAs and reinsurers that do their own underwriting. Headquartered in San Francisco with an office in London.

gobraven.com

INTRODUCTORY GUIDE · UPDATED AUGUST 2026 · REVIEWED QUARTERLY