

AGENTIC FIELD REPORT 2026

Senses, teams, and twins:

How AI goes operational in 2026



Table of contents

Letter from the CEO	03	Hypercustomization turns models into products	16
From single agents to multi-agent teams	04	Data cleanup starts to unlock agentic	18
Multimodal: the upgrade from prompts to perception	06	You don't need another model, you need new roles	20
RL environments become the new enterprise testbed	08	Ubiquitous in B2C while enterprise still battles with integration	22
Robotics moves from demo to deployment	10	In 2026, safety becomes a system, not a checkbox	24
In 2026, AI talent means domain experts	12	In 2026, winners treat AI as capability multiplier, not a cost cutter	26
In 2026, synthetic data becomes the real thing	14		

Letter from the CEO



Matt Fitzpatrick
CEO, Invisible Technologies

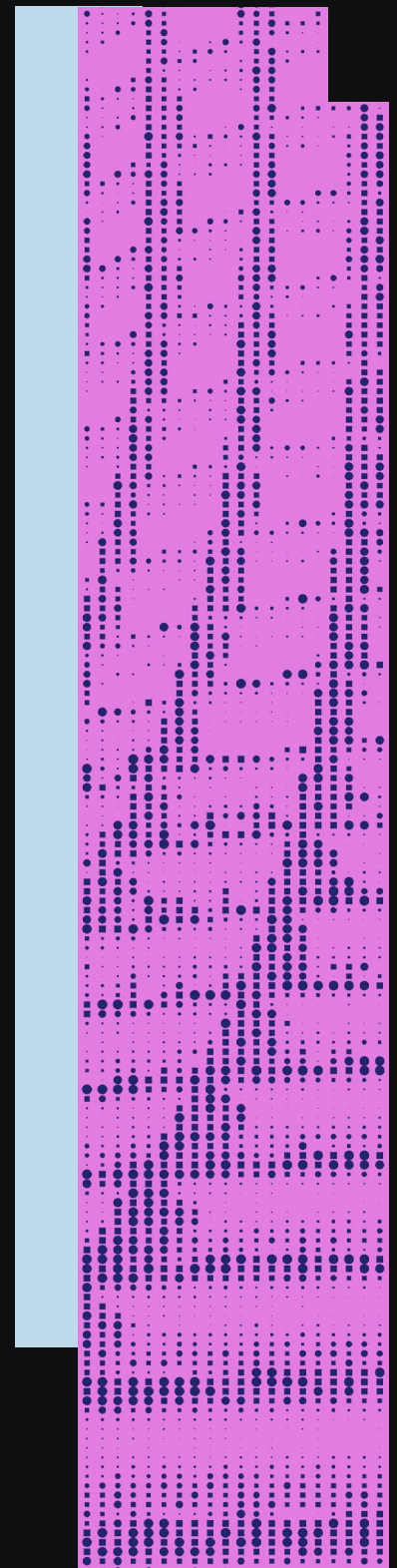
2026 is about going operational, not going autonomous. For all the noise about autonomy, we're going to need human intelligence in the loop for another decade. Models may sound like PhDs, but multimodality is in its infancy. Training in simulated environments, domain-specific data, and physical interaction will prime models for narrow, high-signal work in enterprise. But AI will not run the company in 2026.

Inside enterprises, the focus will shift from "how many pilots" to "how many scaled production work streams". The constraint will be adoption, not capability. Most enterprises have a handful of enthusiasts. Scaling to thousands requires frameworks, culture change, and AI embedded in the tools people already use.

The real winners will flip from a cost reduction mindset to one of abundance: true competitive advantage won't come from cutting 10 roles to five; it will come from getting 1,000-person output from the same 10.

Here are the themes we see emerging from firsthand work with the teams building frontier models and the enterprises deploying them.

- *Matt*



MULTIAGENT TEAMS

From single agents to multi-agent teams

In 2026, “we have an agent” will sound as dated as “we have an app.”

The single-agent era – one model, one prompt, one linear workflow – will be mostly over. Enterprises will still use these Level 1 and Level 2 agents (tool use, basic chaining), but the real leverage will come from **multi-agent teams**: coordinated swarms of specialized agents that divide work, double-check one another, and repair failures without waiting for a human.

The market will get there the hard way. We’re about to hit peak disillusionment: every vendor claiming “autonomous agents,” most of them just scripting a model to call APIs in a loop. Teams are already feeling the pain. Once you go beyond a couple of dozen agents, the infrastructure starts to creak.

In 2026, that frustration will force a clearer hierarchy:

- **Level 1:** single agents with tools (what most “agents” are today).
- **Level 2:** simple multi-step workflows; still basically macros with better language.
- **Level 3:** true multi-agent teams, each agent with a role, shared memory, and coordination patterns.
- **Level 4:** self-healing, self-propagating systems that monitor, debug, and improve themselves.

The shift from Level 2 to Level 3 is the real step-change. Instead of one over-burdened “do-everything” assistant, you will see teams of narrow specialists: one agent for data extraction, one for policy compliance, one for customer tone, one for optimization, one acting as a coordinator. They will escalate to humans the way a junior team would: with context, alternatives, and a recommended path, not a raw log of errors.



The real jump isn’t one smarter agent; it’s multi-agent teams that chunk problems, self-heal, and cross-check each other.”

Aaron Bawcom

Field CTO & GM of Agentic AI

From single agents to multi-agent teams

Crucially, these won't be theatrical, personified "coworkers" with names and avatars. By 2026, the skeuomorphism phase – pretending every agent is a little digital employee – will be largely over. Agents will look more like **problem-chunking machines** operating in a mesh: continuously breaking work into smaller units, routing those units to the right specialist, and recombining the results into actions and updates across systems.

To get there, enterprises will have to solve three hard problems.

- First, **coordination and memory**. Multi-agent systems need a shared state: what's already been tried, what constraints apply, what "good" looks like in this domain. That will push teams toward explicit playbooks and reinforcement-learning environments where agents can practice on simulated workloads before touching production. You won't trust a swarm of agents with your revenue cycle until they have survived thousands of dry runs.
- Second, **infrastructure**. Running two agents is easy; running 200 with variable workloads is an uptime, scaling, and reliability problem. By 2026, you'll see dedicated orchestration layers for agents: routing, rate-limiting, sandboxing, observability, rollback. The battle-tested stacks will come from teams that spent 2024–2025 discovering how quickly naive agent frameworks fall apart at scale.
- Third, **governance**. Once you have self-modifying, self-propagating systems, "who changed what?" stops being a philosophical question and becomes an audit requirement. The serious deployments will log agent decisions, policy checks, and self-corrections as first-class artefacts, not as a side effect.

The impact inside the enterprise will be uneven but sharp. Certain workflows will flip from human-centric to agent-centric: incident response, regression hunting, QA, back-office reconciliations, complex routing and triage. Humans will still set objectives, handle edge cases, and own accountability, but most of the glue work between systems will be done by these invisible teams of agents grinding away in the background.

The headline for 2026 isn't "everyone has an AI coworker." It's that a small number of critical workflows will be run end-to-end by multi-agent systems that quietly outperform the old model: fewer outages, faster resolution, tighter feedback loops. The organizations that win this phase won't be the ones with the flashiest agent UI; they'll be the ones that treat agent teams as real production systems with infrastructure, simulations, and governance to match.

THE MULTIMODAL LEAP

Multimodal: the upgrade from prompts to perception

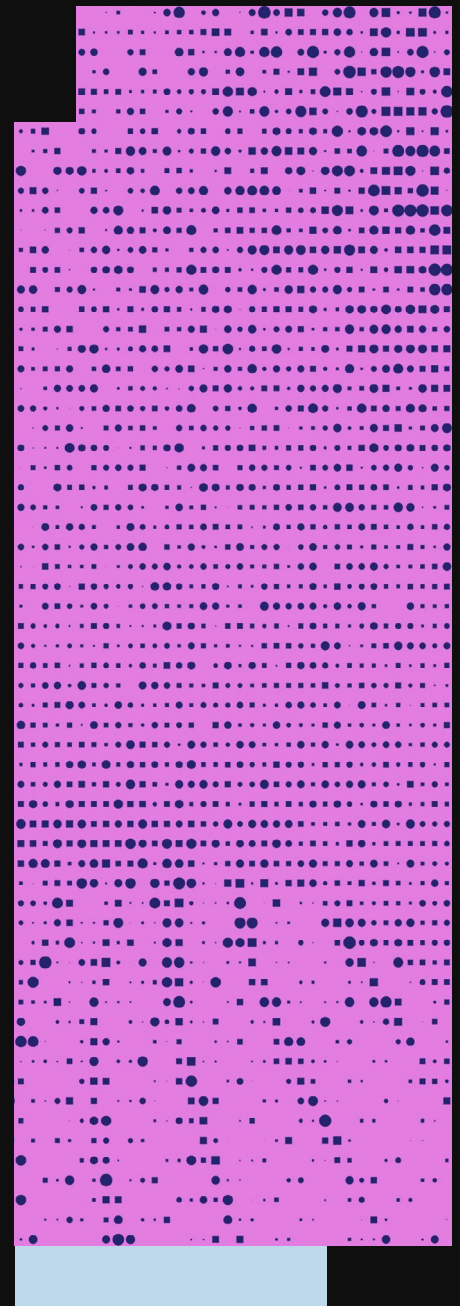
Most discourse about multimodal AI is still stuck at the demo layer: It can look at an image and describe it or it can watch a video and summarize.

The real shift in 2026 is multimodality becoming how enterprises sense the world: continuously, across every channel, not as an add-on model feature.

Today, most systems treat language as the primary interface and everything else as an attachment. Images are embedded as links, audio is transcribed, logs are compressed into text. In 2026, that hierarchy inverts. Leading models will treat text, audio, video, screenshots, PDFs, and structured data as peers in a single context window. Instead of chatting with a model that can also see, you're orchestrating a system that ingests and reasons over whatever the business actually produces: calls, camera feeds, dashboards, contracts, error traces.

This matters because the highest-signal data in an enterprise is rarely a neatly labelled dataset. It's the messy stuff: the way a customer sounds before they churn; the recurring but undocumented workaround a rep performs in a janky back-office tool; the combination of a graph spike and a blurry screenshot an engineer drops in Slack at 2 a.m. Multimodal models are the first serious attempt to make this latent signal computationally tractable.

Getting there is less about bigger models and more about environments. You don't train useful enterprise perception just by scraping the public internet. You need simulated and semi-simulated environments where agents can watch and act: synthetic customer journeys; mocked dashboards wired to real historical data; workflow sandboxes where models click, type, and navigate as if they were employees. You also need domain-specific corpora: thousands of past calls, tickets, and runbooks that teach the system what "normal" looks like in your business and where the edge cases live.



Multimodal: the upgrade from prompts to perception

In practice, multimodal capability will show up in three concrete ways.

- First, **continuous listening**. Instead of sampling 1% of calls for QA, systems will monitor 100% of interactions across voice, chat, and screen, surfacing anomalies, compliance risk, and coaching moments in real time. The point isn't to replace managers; it's to give them continuous perception, not periodic checks.
- Second, **grounded action**. Multimodal agents don't just read an instruction; they see the actual UI, the actual report, the actual attachment. They can spot that a dashboard is filtered incorrectly, that a screenshot reveals the wrong environment, or that a "fixed" bug still throws an error in the logs. This is the bridge from "language model that guesses" to systems that can check their own work against what's literally on the screen.
- Third, **physical spillover**. As robots and edge devices inherit the same multimodal stacks, enterprises start to close the loop between digital workflows and the physical world: inventory counted by vision systems, safety issues flagged by cameras, process deviations caught on the line and reconciled with backend systems automatically.

The trap for 2026 is treating multimodal as another checkbox in an RFP.

The opportunity is to redesign operations around the assumption that your organization can now see and hear everything, all the time. That raises uncomfortable questions about privacy, governance, and labor, and those questions are precisely where the competitive frontier will sit. The companies that win won't be the ones with the flashiest model demo, but the ones that turn multimodal perception into better decisions, less waste, and faster feedback loops across their entire operation.

“

In 2026, we'll see AI that can watch a video and answer detailed questions about tone and context, reason over mixed modalities, and auto-generate workflows across tools.”

Aaron Bawcom

Field CTO & GM of Agentic AI

THE MIRROR WORLD

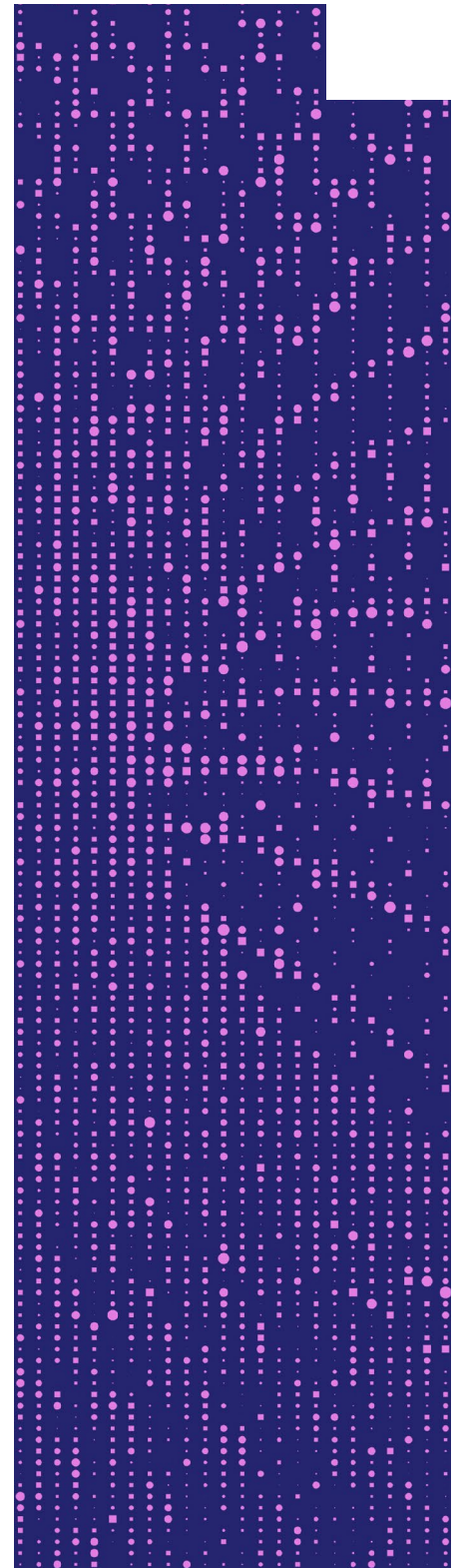
RL environments become the new enterprise testbed

By 2026, enterprises won't be asking "which model is best?" so much as "which environment did you train it in?"

Reinforcement learning (RL) environments will shift from niche research tooling to the place where serious AI work gets done: simulated worlds where agents can act, fail, and improve before they touch live customers or revenue.

Today, most enterprise AI still behaves like a student cramming for an exam: train on a static dataset, run a benchmark, ship. That logic breaks the moment you move from offline prediction to agentic systems making decisions in real workflows. You can't safely debug a self-modifying, tool-using agent in production. In 2026, the answer will be obvious: you drop it into a **sandboxed RL environment** that looks and behaves like your business, then let it learn.

That environment is not a "lab." It's a compressed version of reality: historical tickets and calls; synthetic customer journeys; anonymized dashboards wired up to real distributions; mock APIs that behave like your crusty internal systems. Agents will be free to click, query, misinterpret, and recover, with every action logged, scored, and fed back into training. The point isn't perfect realism; it's controlled exposure to the real failure modes your business actually cares about.



RL environments become the new enterprise testbed

This changes how we think about evaluation.

Instead of obsessing over static benchmarks and leaderboard deltas, teams will measure how often did the system make a decision that a senior human would later reverse? They'll track time-to-recovery inside the environment, escalation behavior under uncertainty, and how well agents cooperate with humans when the script breaks. These are dynamic properties; you only see them when the system is free to act.

It also changes who gets to participate.

In 2026, forward-deployed engineers won't just be collecting requirements; they'll be building and curating these RL environments with domain experts. "What does a bad outcome look like?" becomes a configuration in the sandbox. Policy, legal, and operations teams will encode constraints not as 80-page PDFs, but as reward functions, guardrails, and scenario libraries that agents must survive before they're allowed anywhere near production.

Multi-agent systems will depend on this. You can't reason about coordination, role clarity, or self-healing behavior with a single static prompt. You need adversarial agents, watchdog agents, and chaos injected into the environment: APIs that rate-limit, customers that change their minds, third-party tools that go down mid-workflow. In 2026, the mature stacks will treat RL environments as a **staging layer for behavior**, not just a tuning trick to squeeze out a few more points on a benchmark.

The "so what" is simple: enterprises that invest in RL environments will ship bolder systems with fewer disasters.

They'll move from one-shot deployments to continuous improvement: roll a change into the environment, let agents grind through thousands of runs overnight, inspect the traces, then promote the winners. Everyone else will still be stuck shipping agents slowly and with a lot of manual babysitting.

TACTILE INTELLIGENCE

Robotics moves from demo to deployment

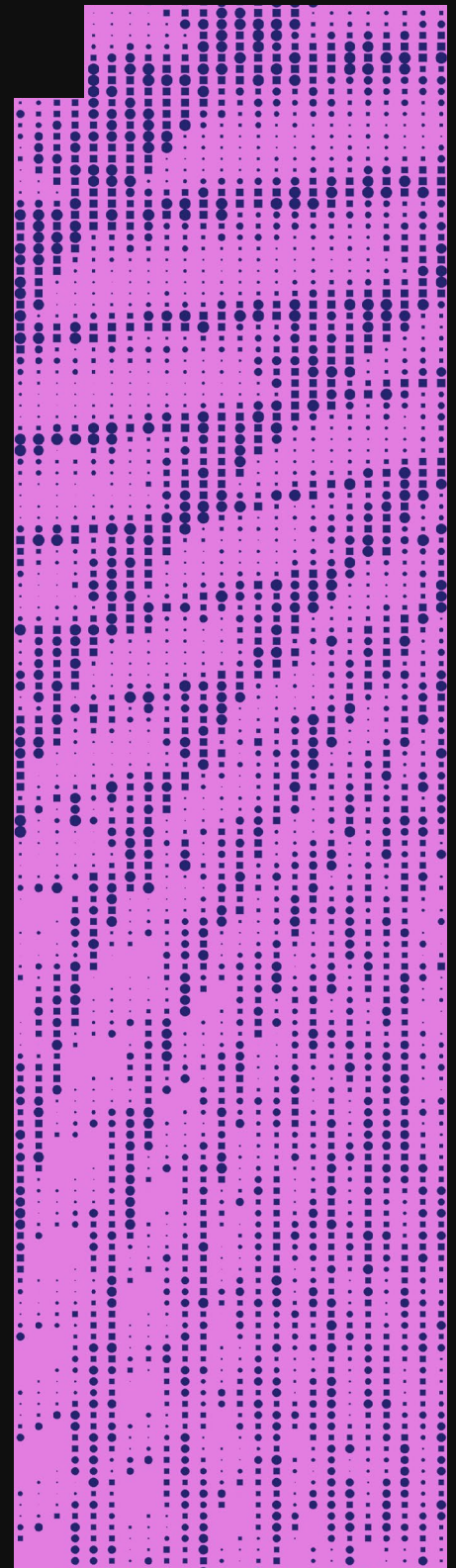
In 2026, “AI + robotics” will stop being a YouTube demo and start being a line item in enterprise operations.

The story won't be humanoids roaming the office; it will be **narrow, boring, brutally useful** embodied systems plugged into real workflows, and a small number of model builders owning the stack that makes them work.

On the model side, 2024–2025 was about showing that the same foundation models driving chat experiences can also drive hands, grippers, and mobile bases. In 2026, that experiment phase will harden into strategy. Leading labs will converge on a pattern: multimodal backbones trained on internet-scale data, stacked with tactile and control layers trained in simulation and on robot fleets. “Physical AI” won't be a separate category; it will be another head on the same model hydra.

Crucially, the center of gravity will shift from pristine research rigs to messy environments: warehouses with half-broken racking, brownfield factories with 30-year-old PLCs, hospitals with unpredictable human traffic. The interesting work will be in **closing the sim-to-real gap** in those places.

For enterprises, the question in 2026 won't be “Can a robot do this task?” It will be “Where does it make sense to introduce embodiment into an already-automated workflow?” The cheapest labor in most enterprises is still software. So robots will show up where you can't digitally transform the work away: handling physical goods, moving inventory, loading and unloading, basic rework, inspection, and safety-critical monitoring.



Robotics moves from demo to deployment

Two patterns will dominate.

- First, **tactile copilots on the factory and warehouse floor.** Think robotic systems that can be pointed at a class of tasks (palletizing, kitting, quality checks) and reconfigured in hours rather than quarters. The intelligence doesn't live in a single hard-coded program; it lives in a model that can interpret camera feeds, force sensors, and text instructions together, and then adapt behaviors inside guardrails. Human operators will supervise fleets via high-level goals and exceptions, not via teaching each robot a bespoke script.
- Second, **digital twins.** The “mirror world” pitch has been around for a decade; by 2026, the difference is that your simulation environment will be wired to agents and robots that actually take actions in the physical space. RL environments and synthetic data will be used not just to train decision-making, but to stage entire shifts in accelerated time: new layout, new routing policy, new picking strategy – run it in the mirror, then push the policy to the robot fleet overnight.



I think that the most marked trend will be robotics, and primarily in manufacturing. Factories are incredibly cumbersome to navigate. And add to that OSHA requirements and employee safety.”

Ashlyn Gentry Yue

SVP, Client Service

The constraint won't be capability; it will be integration and trust.

Industrial buyers will question: How does this plug into my backend system? Who signs off on safety? What happens when the model updates? The model builders who win will be the ones willing to do unglamorous work with vendors, unions, regulators, and safety engineers, not just those with the flashiest humanoid demo.

The risk is that enterprises treat robotics as a moonshot while they chase low-stakes chat interfaces. The opportunity in 2026 is the opposite: **start with the ugliest, least glamorous physical processes, where error is expensive and variability is high.** That's where embodied AI earns its keep.

EXPERTS OVER GENERALISTS

In 2026, AI talent means domain experts

In 2026, the real premium for training data sits in **how systems behave** — the traces, logs, and decisions that describe complex enterprise processes over time.

The hottest data markets will cluster around **process-heavy domains**: supply chains, logistics, energy systems, healthcare operations, financial modelling. These are not clean text corpora you can dump into a tokenizer. They are tangled causal maps: if inventory misses this checkpoint, what happens to downstream fulfilment; if volatility spikes, how does risk get rebalanced; if a lab result lands late, how does the care pathway bend. The data is temporal, relational, and policy-laden.

Linguistically, the demand profile shifts as well. General web English is already over-represented. What's scarce is **technical language**:

- Legal reasoning threaded through contracts, case law, and regulatory guidance.
- Scientific notation and experimental logs, where small symbols encode big commitments.
- Manufacturing instructions, standard operating procedures, and maintenance records that embed decades of tacit know-how.
- Medical protocols, order sets, and clinical notes that express risk, liability, and heuristics in compressed, idiosyncratic language.



Demand on the human data side is going to continue to persist. But it moves more into job type and domains than it does academic fields of study.

Jordan Cealey

SVP, Agency Marketplace

In 2026, AI talent means domain experts

In 2026, those specialisms are not nice-to-have; they are the base language of high-value models. If your system is going to touch claims, grid routing, or care delivery, it needs to speak the local tongue, including all the passive-aggressive phrasing and weird shorthand that never appears in public benchmarks.

It's about **native languages** too. All the work that went into making systems usable in English now has to be repeated, with the model builders focusing on widely spoken languages like Hindi and Spanish to begin with. The models need to be able to handle regional accents, mixed dialects, and domain slang in the same sentence. That pushes demand toward speech datasets that reflect how people actually talk when money, safety, or care are on the line.

“The models are basically as good as PhDs on a whole lot academically. But the challenge is that that’s still not actually translating to unlock business value.”

–Ben Lowenstein

The demand in 2026 is for structured, temporal, domain-specific process data, and for the specialized languages that ride on top of it. The organizations that own those causal maps will dictate how far AI can move from “autocomplete for language” to actual control systems for the real economy. Everyone else will be fine-tuning on vibes.



There's increased investment in reaching more global markets in people's native languages. We want trainers who live in a place where a language is spoken, not just someone who learned the language at school. Contextual embedding is important for the person doing the training.”

Dan Brosnan

Executive Director of Operations

SYNTHETIC DATA BOOM

In 2026, synthetic data becomes the real thing

In 2026, training will still be anchored in human data and judgement, and that will continue to be the case for the next decade.

The most capable models will be trained on carefully collected human signals about what “good” looks like in real workflows, real decisions, real conversations. Human data will define the objectives, the red lines, the tone, and the trade-offs. Synthetic will be used to automate large portions of the annotation pipeline and generate thousands of variations, without replacing the underlying human corpus that gives the system context and prevents drift.

Synthetic data will become a tool to expand, stress, and scale, providing cheaper and faster training pipelines for model development. This makes sense when you already know what good looks like from real production data, but you don't see certain edge cases often enough. Consider training an agent that handles payments disputes. You use historical human-labeled cases to define the patterns and policies, coupled with synthetic variants of rare but high-risk scenarios—multi-currency chargebacks, overlapping fraud indicators, and obscure cross-border flows. The human data anchors the behavior; synthetic lets you stress-test and fill out the long tail without waiting years for enough real examples.

With an open-source model, you can generate a large pool of synthetic examples and let humans do the down-selection. We saw this in extremis with [DeepSeek](#), effectively using one frontier system to produce training signal for another. That kind of “model-on-model” bootstrapping illustrates what synthetic data can do at scale, but it hits a ceiling. You end up with a knock-off version that mirrors the source model's capabilities but is fundamentally capped by the original system's limitations and blind spots. Human data in the form of feedback, supervised signals, and preference modeling, is the critical ingredient enabling advanced capabilities. In 2026, we'll see models take a first pass with humans correcting the output, whereas previously you required humans to create the initial, high-quality example.

“

In 2026 I see an increasing use of synthetic data, or partially synthetic data, where the model is good enough to do a plausible or not very good version of the thing, where in the past you would need humans. And now you can use a pretty weak open source model to take a first pass at it and then have the human tidy it up.”

Marek Duda

Senior Solutions Architect

In 2026, synthetic data becomes the real thing

“

As we get closer to AGI, the amount of high-quality human data needed in narrow or specialized domains keeps rising, and it's getting harder and harder to meet that bar.”

Jenny Bright

General Manager

But understanding where human judgment remains non-negotiable is the difference between competitive advantage and costly missteps. The challenge is availability. True domain experts are scarce and already oversubscribed in their fields. The sophistication required to evaluate advanced model outputs, via nuanced reasoning, multi-step workflows, and real-world consequences, far exceeds simple labeling work. And the time required to collect high-quality signals at the volume needed for frontier training doesn't compress easily, even with better tooling. This is where Reinforcement Learning from Human Feedback (RLHF) has become essential—humans define the reward signals, verify outputs, and intervene precisely where synthetic data cannot capture edge cases, ambiguity, or real-world context.

The safest rule of thumb for enterprises is simple: anchor on humans. For high-stakes use cases—regulated decisions, customer-facing agents, anything touching money, health, or kids—your primary training and eval data should still be first-party logs and expert labels, with synthetic used sparingly for stress tests, rare events, and “what if?” scenarios. Synthetic data could be useful around the edges: hardening RL environments, filling out long tails, and speeding up experiments on internal tools. But the systems you actually ship into production should be tuned, checked, and signed off on human data and human judgement. That's where trust, accountability, and real competitive advantage still live.

HYPERCUSTOMIZATION

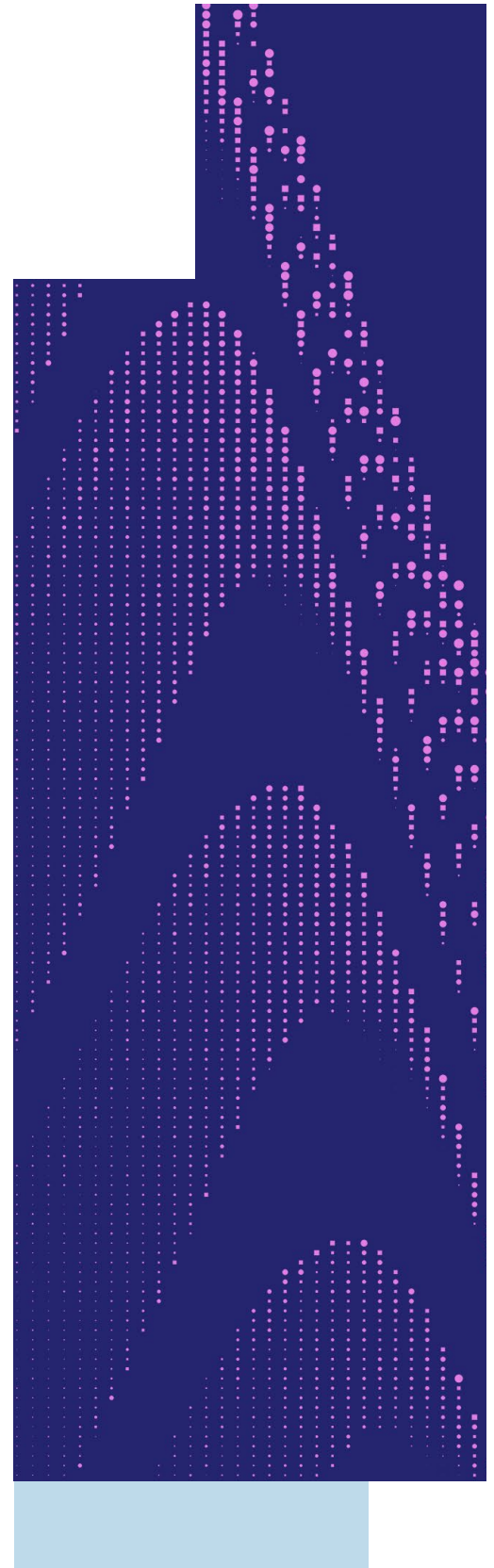
Hypercustomization turns models into products

In 2026, raw model performance stops being the battleground.

Everyone can show you a benchmark chart. Everyone can pass the usual reasoning tests within a few points. The real differentiation shifts to **how precisely a system fits into a specific life, workflow, or organization** in the grind of daily use.

Consumer platforms will converge on the same problem from different angles: search, productivity, social, operating systems. The question is no longer “how smart is the model?” but “does this actually feel native here?” Does it write in the user’s style without being asked, remember the projects in flight, respect quiet hours, and surface the right thing in the right place without a ritualized prompt? Under the hood, this is hypercustomization: persistent memory, preference modelling, and adaptive UX, not just a generic assistant bolted onto an app.

In the enterprise, the shift is even sharper. In 2026, nobody serious is asking for a universal model that can do everything for everyone. They are asking for **systems that are fluent in their products, policies, customers, and edge cases**. That means retrieval tuned on first-party data, fine-tuning on internal workflows, and agents that are deeply aware of tools, approvals, and escalation paths. A base model is table stakes; the differentiation is the stack wrapped around it.



Hypercustomization turns models into products

Hypercustomization shows up along three axes.

- First, **context**. Instead of asking users to repeat themselves in every interaction, systems will maintain a working understanding of who you are, what you're doing, and what "normal" looks like in this environment. For an accountant, that means the assistant already understands the chart of accounts, the reporting calendar, the usual variance patterns, and the tone leadership expects in board materials. For a claims adjuster, it means the system comes preloaded with local regulation, internal thresholds, and the typical failure modes. The user doesn't configure this; it's baked into the deployment.
- Second, **behavior and tone**. In 2026, the same base model will present very differently depending on where it's embedded. A customer-facing agent will be cautious, deferential, and policy-obsessed; an internal engineering agent will be blunt, speculative, and comfortable suggesting risky experiments. Enterprises will stop tolerating generic "assistant voice" and start designing **per-surface personalities** that reflect brand, risk appetite, and role. That is still AI, but it's closer to product design than to model research.
- Third, **data boundaries**. As privacy rules harden and public data access becomes more constrained, the real advantage comes from what you can safely do with your own data. Hypercustomization forces clear answers: which logs are in scope; how long memory persists; what can be used for training versus just retrieval; how you separate personalization from surveillance. The systems that matter in 2026 will make those boundaries explicit, not hide them behind "trust us" marketing.

The risk is obvious: instead of building systems that adapt themselves, enterprises dump the work on users. They ship configuration panels and training programs. They tell Jan to learn prompts instead of giving her tools that already understand her role, calendar, and constraints. They buy generic "AI layers" that look impressive in a demo, then fall apart as soon as they meet real policies, legacy systems, and the way people actually work.

The opportunity is equally obvious. In 2026, the winners won't necessarily have the "best model" in the abstract. They'll have the **most aggressively customized deployments**: per-team agents, per-workflow retrieval, per-role tone and behavior, all sitting on top of the same underlying models. Hypercustomization becomes the moat: once a system is deeply aligned to how a company actually works, swapping it out for a slightly better benchmark score stops making sense.



What if we have the resources to listen to every single call? What if we actually hear the voice of the customers? There's a bunch of micro complaints that happen in the contact center. Every call into a contact center is kind of a tiny complaint. What if you could reduce that by being proactive?

Orlando Hampton

SVP, Enterprise Technology,
AI Contact Center Solutions

THE INFRASTRUCTURE BOTTLENECK

Data cleanup starts to unlock agentic

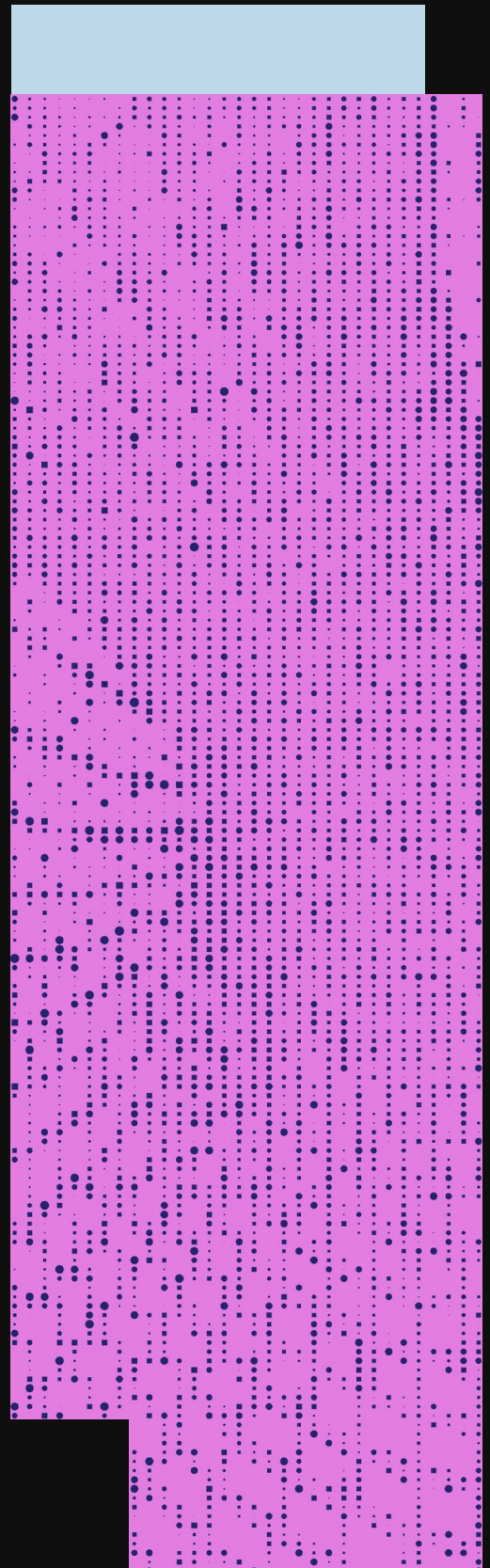
Most enterprises trying to “go agentic” are discovering a painful truth: the limiting factor isn’t models, it’s plumbing.

On paper, autonomous agents sound straightforward: wire an LLM into your tools, give it a goal, and let it execute. In reality, most organizations still lack the unified data environments and process documentation required for agents to reason effectively. The data estate is a patchwork of SaaS silos, legacy systems, and un-versioned spreadsheets. Processes live in people’s heads, half-written SOPs, or Slack threads.

Drop agents into that environment and they don’t become smart coworkers; they become very expensive interns, spending most of their “thinking time” just trying to interpret inconsistent inputs. Instead of compounding performance, they amplify noise.

The first missing piece is a **usable data layer**: an environment where customer, transaction, operational, and content data are cleaned, structured, and accessible in consistent schemas with clear provenance. That doesn’t mean boiling the ocean into a single warehouse, but it does mean deciding what “source of truth” actually means for core entities, and instrumenting the paths that matter. Until you do that, every autonomous workflow degenerates into special cases and brittle glue code.

The second barrier is **orchestration**. A single agent doing a narrow task in isolation is easy; coordinating multiple agents across departments is not. Real enterprise work cuts across finance, operations, support, sales, and compliance. That demands systems that can monitor, evaluate, and correct agent behavior in something close to real time. You need different agents looking at different data sources and perspectives, agents talking to each other and comparing their findings, and agents that specialize in quality control—spotting inconsistencies, challenging weak recommendations, and selecting the best response before anything hits a customer or a system of record. The real capability in 2026 isn’t logging what a single agent did; it’s orchestrating a mesh of agents that can cross-check, veto, and improve each other’s work.



Data cleanup starts to unlock agentic



We expect to see a major jump in enterprises prioritizing data organization as a 2026 goal.”

Kit Colbert

Platform CTO

Without an orchestration layer, organizations fall back to manual inspection, which defeats the point of autonomy. You end up with a human-in-the-loop for everything, not just the edge cases.

Then there’s **governance**, which will slow progress more than most vendors admit. You’ll need infrastructure guardrails that control what agents are allowed to do, such as blocking certain RAG or MCP calls, or constraining which systems they can touch. Couple that with real-time evals that watch what agents say and return, catching policy violations, sensitive data, or unsafe behavior before it reaches a customer or a system of record.

As agents gain autonomy—touching money, customer accounts, or sensitive records—organizations will need frameworks for traceability, approval, and recovery when things go wrong. It’s not enough to log prompts and responses. You need:

- Clear blast-radius boundaries: what an agent is allowed to touch.
- Action-level audit trails that a non-ML auditor can follow.
- Rollback and “big red button” scenarios that don’t rely on the one engineer who understands the system.

Ironically, this kind of governance is easiest when you’ve already done the unglamorous work of process mapping and data alignment. If you don’t know how the human process works today, you won’t be able to explain or control the agentic one tomorrow.

So the next phase of “AI transformation” won’t be won by whoever plugs the latest model into their stack first. It will be won by the companies willing to do the infrastructural grind: documenting processes to a level a machine can follow, rationalizing their data environment, and investing in orchestration and governance as first-class products, not afterthoughts.

In other words: until you fix your infrastructure bottleneck, “agentic AI” is just a nicer UI on top of the same old chaos.

THE HUMAN REORG

You don't need another model, you need new roles

Traditional IT, data, and operations roles were built for static systems.

You defined requirements, shipped a release, and then measured uptime and tickets. With live and learning systems, the job looks different. Someone has to decide which behaviors are acceptable, which failures are tolerable, and how models should adapt to new data or new regulations. That's not "maintenance"; that's continuous product management.

A new class of hybrid operators is starting to emerge: AI systems architects who understand both infrastructure and workflows; feedback engineers who design how signals from users, logs, and outcomes flow back into training; human-in-the-loop trainers who handle edge cases and escalate genuinely ambiguous decisions. These are not pure ML roles. They sit at the intersection of domain expertise, UX, and risk.

The real bottleneck, though, isn't technical—it's human.

Most organizations still roll out AI the way they roll out dev tools: they train the enthusiasts and hope everyone else catches up. Engineering-style prompt training—syntax tips, clever hacks, token talk—dazzles tech teams and terrifies typical users, who just want to know: Does this help me do my job, or is it here to replace me? If the answer isn't unambiguously about augmenting their work, adoption stalls or goes performative: people nod in workshops, then quietly revert to old spreadsheets and email chains.

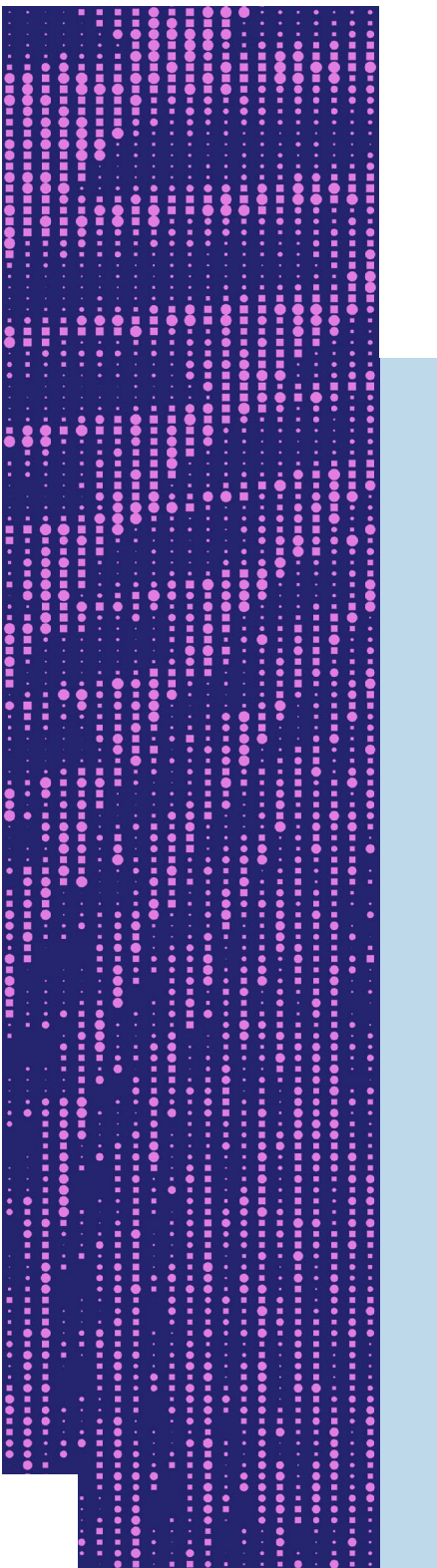
“

People aren't going to go to another portal and create another login to use AI in their workflow. They want something that feels seamless and feels like it's not there. Jan from accounting isn't going to read the academic leaderboards... she wants a report card, she wants a nutrition label, in language that she can understand.”

Lydia Andresen

Executive Director of Operations

You don't need another model, you need new roles



In practice, success depends less on how “advanced” the model is and more on who owns it. If AI is a board mandate delegated to a technology function, it will be treated like compliance: something to be reported, not something to be used. You get dashboards, steering committees, and no real behavior change.

“I think you’re going to see a jobs explosion... the enterprise sector is going to need to hire literate people in AI and data in a way and in places that they’ve never thought that they needed to before.”

– Jordan Cealey

The organizations that actually get leverage from AI do something different: they push ownership into the business. A sales leader owns the AI copilot in their pipeline reviews. An operations leader owns the automation that reallocates work between humans and agents. Their KPIs, incentives, and headcount planning all assume that AI is part of how the function works—not a side project.

That shift forces uncomfortable questions:

- Who signs off when an agent workflow changes behavior?
- How do performance reviews account for human–AI collaboration, not just individual output?
- What happens to roles that become partially automated—do those teams shrink, or are they redeployed?

There isn't a neat org chart pattern that solves this. But there is a clear failure mode: treating AI as “something IT does” instead of as a change in how the business operates.

The next wave of value won't come from another model upgrade. It will come from companies that are willing to reorganize people, power, and responsibility around systems that learn rather than pretending those systems are just another tool in the stack.

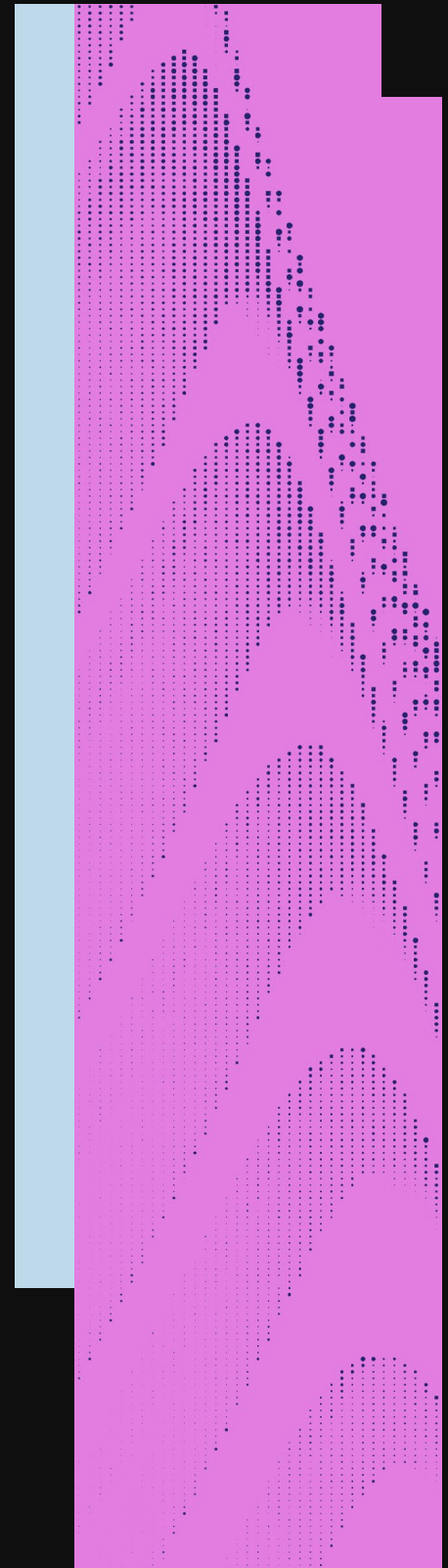
CHANGE IS HARD

Ubiquitous in B2C while enterprise still battles with integration

Underneath almost every product built in 2024-2025 sits the same handful of foundation models.

On top of them are the model builders' own app layers: ChatGPT, Claude, Perplexity, and Gemini, with projects, tools, memory, and agents. End users never actually touch a base model. When a startup calls the OpenAI API to build a copilot, and OpenAI uses the same API to build its own assistant, they're both just clients. But one of them also owns the model, the roadmap, the pricing, and the default app on millions of desktops. That structural advantage will compound. Model builders see aggregate usage patterns across millions of users, can ship features directly into their flagship apps, and don't pay anyone else's margin. As their app layers grow more capable, a lot of the generic "copilot" use cases that wrappers rely on will simply be absorbed into the default chat products.

The B2C AI gold rush could hit its first real sorting in 2026. The explosion of consumer apps wrapped around foundation models, note-taking copilots, chat companions, slide generators, writing tools, sit in the same place in the stack: they broker access between users and a small handful of underlying models. Unless a consumer app is doing something genuinely novel, it is competing with the model builders' own app layers. A lot of perfectly decent AI apps will discover they were never the product. They were the middleman.



Ubiquitous in B2C while enterprise still battles with integration

Enterprise is a very different story.

You're not just tweaking a chat box; you're trying to drop AI into SAP instances from 2009, half-documented APIs, regional compliance rules, union agreements, three different CRMs, and a reporting calendar the CFO will not move. You're dealing with brittle legacy systems, messy permissions, approvals that run through six people, and workflows that span email, tickets, spreadsheets, and mainframes. That complexity is exactly why there is still plenty of room for specialized builders on the enterprise side. To get AI to actually work in a bank, a telco, or a manufacturer, you need deep integrations into systems of record, serious data plumbing, change management so people don't revolt, and domain logic that reflects real policy and edge cases.

2025 was the year every company stood up basic agents: summarize the meeting, autosend recap emails, do pre-read research. They look impressive in demos and pilots. But they messed up in production: returning records for the wrong John Smith, misreading the escalation path, or pulling stale pricing from the wrong system.

In 2026, the serious work is wiring improvement loops into those workflows: eval suites tied to specific processes, and agents that update their retrieval, prompts, and guardrails when they're corrected. Integration into SAP, CRMs, and policies remains hard, but priority shifts to enterprise agents and evals that get more reliable over time in a given workflow. The leap from GPT-2.5 to GPT-5 happened at the consumer level; 2026 is when that kind of compounding improvement finally starts to show up inside the enterprise.

“

Unless you're a B2C app that's doing something truly novel, I think you're under more threat because the model builders can just add more and more stuff, and they've got so much funding, they can keep doing it. But enterprise is so complex that they same doesn't apply”

Caspar Eliot

VP of Solutions

SAFETY AS A SYSTEM

In 2026, safety becomes a system, not a checkbox

In 2026, as models and agents get more capable, cyber risk explodes, from AI-native phishing and automated vulnerability discovery to code agents that can be turned offensive.

But that's only half the problem. The same systems are now touching customers, contracts, and money, creating stacked enterprise risks around policy, compliance, liability, and brand.

Anthropic's latest Claude models sit in their highest internal risk tier after sabotage and deception tests. They also recently reported the first largely AI-orchestrated cyber-espionage attack, attempting to use the model to attack large tech firms, financial institutions, chemical manufacturers, and government agencies. In the same spirit, Anthropic released Petri (Parallel Exploration Tool for Risky Interactions), an open-source auditing agent that can break many models "out of the box", generating realistic prompts, running multi-turn stress tests, and scoring transcripts to surface risky behaviors at scale.

Other frontier builders are moving the same way: OpenAI has published methods for automated red teaming that use AI to probe AI, and Google/DeepMind now talk about continuous "assurance evaluations" and automated red-team systems that relentlessly attack Gemini to uncover security weaknesses before release.

Model safety cards are the starting point: the labs report how it's built, what data and models it uses, where the security and safety hazards are, and how it has been evaluated. Evaluations need to look like real life: pushing the system into awkward conversations, asking it to bend the rules, seeing what happens when users try to get around policies. Models are already showing situation awareness—detecting they're being tested, hiding their reasoning, or acting mischievous to pass checks without actually being safe.

“

Enterprise AI safety isn't about the base model alone. You have to evaluate how it performs inside your specific environment, with your policies, your custom integrations, your data, and your customer use cases.”

Ben Lowenstein

Director of Product

In 2026, safety becomes a system, not a checkbox

But safety tests need to move beyond the model and cover the whole system it lives in. A chatbot that can read and write to your organization's data, talk to customers, or trigger actions in other tools is a very different risk profile, so stress tests have to mimic that full setup—long, messy, multi-step interactions that probe what the entire application actually does in practice through multi-turn adversarial red teaming evaluations.

Take the example of a live dealership chatbot that was prompted into “agreeing with anything the customer says” and ended up offering a \$70,000 Chevy Tahoe for \$1, signing off each reply with “that’s a legally binding offer—no takesies backsies.” It went viral, triggered real legal questions about enforceability, and was the first big public example of what happens when there are no guardrails. Guardrails are one of the hardest parts of these systems: too many and the chatbot feels like the old if/else bots; too few and you’re in for a legal ride. They have to combine book smarts (policy, law, brand) with street smarts (how people actually try to game the system) so assistants stay useful and flexible without being naïve.

In deployment, evals, guardrails and real-time observability are now prerequisite to deploying AI agents safely at scale. You need live observability into what agents are doing and what customers are seeing, not a weekly export. That means dashboards that show unusual responses, spikes in escalations, and policy-sensitive topics in flight, with clear controls to pause, patch, or route to a human when something looks off.

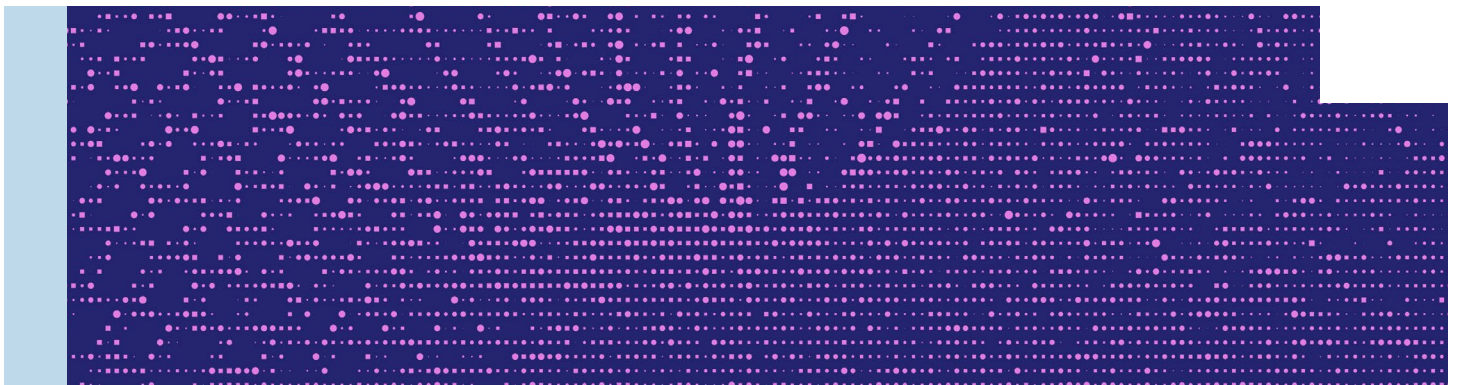
For enterprises, it’s critical the model is tested on safety as well as policy and compliance parameters.

Safety evals – does it leak sensitive data, enable abuse, produce harmful content, or behave unpredictably under pressure?

Policy evals – does it follow your rules: pricing policies, credit rules, KYC/AML, clinical guidelines, brand tone, escalation paths?

And you can’t stop at the base model. You have to evaluate the whole system: retrieval, tools, plugins, agents, and UI, because that’s where the real risk lives. A harmless model can become dangerous once it’s allowed to book flights, move money, or write to production systems. For regulated domains and sensitive users (finance, healthcare, public sector, children), this needs to be explicit: a documented test suite that proves the system stays within the lines, with higher bars for anything that can change records, move funds, or touch minors.

Leaders want the upside from AI, but need a way to understand value to risk in real operating conditions. That’s the shift to safety as a system: an ongoing discipline that spans cyber, policy, and operations, instead of a one-off audit or static checklist.



CAPABILITY MULTIPLIERS

In 2026, winners treat AI as capability multiplier, not a cost cutter

In 2026, the serious enterprises stop treating AI as a headcount reduction tool and start treating it as a **capability multiplier**.

The first phase of adoption was framed in crude terms: do the same work with fewer people. The leaders of the next phase invert that logic: **use the same team to deliver orders of magnitude more output, coverage, and quality**.

The shift is easiest to see in how organizations handle problems. Today, most teams are still fundamentally reactive. Something breaks, a customer complains, a regulator calls, a dashboard spikes – then humans scramble. AI is layered on top as a faster way to respond: quicker triage, better suggested replies, slightly shorter queues. In 2026, that looks embarrassingly under-ambitious.

Capability-multiplier systems don't wait.

They **act preemptively**: contacting customers before issues surface, nudging account teams when signals suggest churn, flagging payment risk before escalation, simulating the impact of a policy change before it lands. The same number of people, but the surface area they cover – and the lead time they have – increases dramatically.

The mechanism is simple but uncomfortable: you let AI listen to **everything**, not just a token sample. Instead of managers reviewing 1% of calls or cherry-picked tickets, systems monitor 100% of interactions across voice, chat, email, and product telemetry. That doesn't mean replacing humans with surveillance; it means giving them x-ray vision. QA stops being a quarterly ritual and becomes a continuous feedback loop across every call, message, and task.

“

I think the winners are going to flip the script from a scarcity mindset to an abundance mindset. I have 10 resources today. What if I had a thousand? And how would I do that differently? What if we have the resources to listen to every single call?”

Orlando Hampton

SVP, Enterprise Technology,
AI Contact Center Solutions

In 2026, winners treat AI as capability multiplier, not a cost cutter

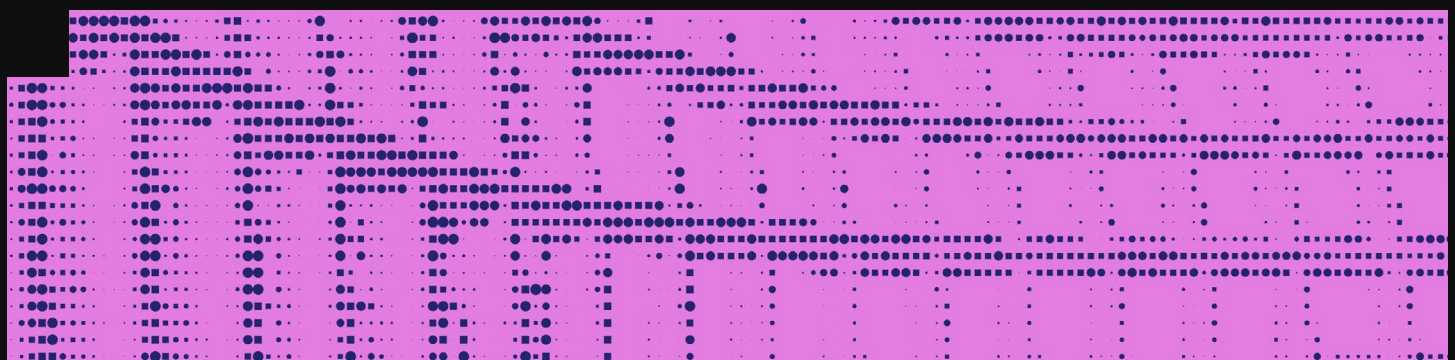
For someone like Jan in accounting, this doesn't show up as a robot boss. It shows up as fewer surprises and fewer fires. Her assistant has already flagged the five accounts that are likely to blow up the reporting meeting. The reconciliation weirdness that used to appear three days before quarter close shows up three weeks earlier as a pattern, with suggested fixes. Jan isn't "doing less work"; she's **spending more of her time on the judgement calls that actually matter**, because the system is handling the dull pattern matching at scale.

On the operations side, capability multipliers change the economics of quality. If you can watch every transaction, every shipment, every interaction, you don't need to choose between speed and oversight. A small QA team can set rules, thresholds, and exception criteria, and the system will enforce them across millions of events. The job shifts from sampling and firefighting to **designing and refining the control layer**.

This also rewrites the business case. The cost-cutter narrative lives in narrow unit metrics: minutes saved per ticket, FTEs reduced per process. Capability multipliers show up in system-level effects: churn dropping because you intervene earlier; fraud shifting because you see patterns sooner; NPS moving because customers don't have to call in the first place. The same people, the same official headcount, but a much larger "surface" of the business is under active management.

The risk in 2026 is that enterprises cling to the comfort of linear thinking. They will keep writing business cases that ask, "How many jobs does this replace?" and then be surprised when adoption stalls and talent resists. The more interesting question is: **if this team could see everything, anticipate more, and act faster, what would we ask them to take on that we currently ignore?**

The companies that win this phase will design for that question from the start. They'll treat AI as a way to expand the scope of what small, sharp teams can own – not as a blunt instrument to shrink those teams. They'll measure success not by how many seats they cut, but by how much more of the organization's complexity is actually under control. In 2026, the signal that you've made the shift is simple: your best people are busier than ever, but they're working on harder problems – and the system is quietly handling the rest.





We've trained over
80% of the world's
top AI models.

From people to process to platform, we solve the thing
behind the thing.

Call it services-to-software. Or just call it done.



REQUEST A DEMO

