Baringa

From chaos to confidence

AI in customer service:
the tech is ready, are you?

Executive summary

Baringa has spent 15 years helping organisations transform their customer service, through the era of digital deflection, chatbot pilots and contact centre consolidation to today, where the pace and complexity of change has never been higher.

This report draws on that experience – and on our recent, in-depth assessment of 17 leading vendors (from a longlist of 60) across 26 criteria, alongside interviews with leaders running these platforms in production. The insight from this independent research is intended to support leaders to make better decisions about AI agent deployment and vendor selection.

Here are the five key findings from our research:

The technology has moved from promise to proof

AI agents are genuinely delivering in production: containment rates of 70–85% for well-defined use cases, cost-to-serve reductions of 20–40%, and individual organisations using AI to manage more than 15 million interactions annually. The debate about whether this technology is ready for enterprise scale is over.

Most pilots stall for reasons that have nothing to do with the technology

The real causes are consistent: ambiguity about what the organisation is trying to achieve, unready data foundations and governance gaps that surface mid-deployment. All are resolvable and all can be addressed before vendor selection. Organisations struggling to realise value after 18 months are not being let down by the technology.

The buying conversation has matured

A year ago, buyers were asking which vendor had the best AI. The mature buyers in our research are asking a different question: which engagement model fits their internal capability, whether their data foundations can support an agent at scale, and what sector-specific accelerators can compress time to value. That shift in question is itself a signal of how far the market has moved.

4 Every AI agent is built on the same ten layers

Many of the underlying components of every AI agent – speech, language models, voice synthesis – are largely commoditised. The layers that create durable advantage are orchestration, agent builder tooling and observability. Most vendor RFPs don't weight them correctly – but these are the ones you should focus on.

5 The decisions you make now are not just about customer service

The architecture you build today is the foundation for customer-owned AI. It will determine whether the infrastructure you have is fit for a fundamentally different commercial relationship with customers. This is a now consideration, not a future one.

Most organisations planning to deploy an AI agent are treating this as a technology procurement decision. The ones achieving results at scale treated it first and foremost as a business transformation decision.

The technology is ready. The question is: are you?

Inside this report

So you want to deploy an AI agent

AI can transform customer service. But great deployments don’t happen by accident.

Millions of customers are enjoying faster, more personalised and predictive service, thanks to AI agents embedded across customer service operations. The technology has matured dramatically over the past two years alone. So, the path to value should be clear by now. For most organisations, it isn’t.

It takes more than a great platform to get great results

The truth is that getting value from AI in service depends on much more than signing up for the latest technology. The value is realised when you get from promising pilots into production. To do this, you need to be clear on what you want to achieve and what’s realistically possible, combined with the knowledge to navigate a noisy marketplace and the best vendor solution for you.

The complexity is real

Setting strategy, scaling up, selecting vendors – these are all things we see leaders wrestling with today. We've spent more than 15 years helping clients transform customer service.

The opportunities now are greater than we've ever seen – but so are the complexities and the risks.

To capture insights from that experience and bring them right up to date, we carried out a new iteration of our AI in service research. We spoke to leaders across sectors, whose organisations are running platforms in scaled production solution, asking them about the challenges they've faced, the lessons they've learned and the results they've achieved. Their journeys inform everything in this report.

Clarity is here

This report gives you the tools to move forward in this space with confidence. Inside, you'll find:

- **Six key questions to answer** before you begin, so everyone's aligned on what to expect and is committed to making it work.
- **Four deployment pathways** that translate your answers into a realistic approach and timeline for moving from pilot into production.
- **An overview of the fast-changing vendor market** and what developments here mean for your procurement decisions and future customer service.
- **A breakdown of the universal ten-layer stack** that underpins every customer-facing agent, to help you make the right calls on what to buy.
- **Insights from our research** so you can learn from others who've already made the mistakes worth avoiding.

If you like what you read here, talk to us about the second part of our research – an in-depth assessment of 17 leading vendors across 26 criteria, looking at what you can expect in terms of capability, fit and commercial arrangements. We believe it's the best in the market.

It's crunch time

In 2026, customers' expectations of service quality are sky high. No organisation serious about meeting them can do this without bringing in AI agents.

The question is no longer whether AI belongs in your service operations. It's whether you're making the right moves to get it there.

This report is designed to help.

Remember when this was just all potential?

Our 2026 research shows AI in customer service delivering real results that are hard to ignore.

The past decade has taken customer service transformation through an era of digital deflection, chatbot pilots and contact centre consolidation.

But the current moment is genuinely different. AI now is enabling organisations to establish a fundamentally different mode of customer interaction and the infrastructure for a different kind commercial relationship with customers altogether.

Two years on from our 2024 research, our conversations with organisations taking part plus a comprehensive market scan shows things have moved on materially.

The technology has moved from promise to proof

Organisations we spoke to reported containment rates as high as 70–85% for well-defined use cases. Cost-to-serve reductions are reaching 20–40%. Individual organisations are automating over 15 million interactions a year. A year ago, the dominant question for enterprise buyers was whether AI agents were ready for production scale. Today, the answer is an unequivocal 'yes'.

Buyers are asking better questions...

Last year, organisations were asking which vendor had the best AI. The mature buyers in our research are now asking different questions: Which engagement model fits our organisation? Do our data and integration foundations support an agent at scale? What industry-specific accelerators can compress time to value. It's a more considered conversation and a more productive one.

...but too many pilots are still stalling

We found that the organisations still running pilots after 18 months aren't being held back by the technology but because they didn't answer key questions before choosing a vendor. Strategic ambiguity, underestimated capability requirements, unready data foundations and governance gaps discovered mid-deployment are the key causes of stalled pilots. They surface after vendor selection and the consequences are the same across every sector. They were true a year ago, and remain true now.

The market is finding its shape...

Our market scan showed market consolidation and convergence accelerating over the past two years. The proliferation of vendors, models, speech engines and orchestrators that defined 2024 and 2025 is giving way to a clearer structure. LLM, speech-to-text and text-to-speech components are increasingly interchangeable. CRM and CCaaS platforms have moved into the orchestration layer, absorbing those components into managed services.

The real differentiation has moved into agent-builder tooling, observability, commercial models and data sovereignty. The specialist AI platforms have focused specifically on those areas. The gap between the leaders is widening and it is now easier to see where it sits.

... but a new risk is hiding in plain sight

Our market scan uncovered a new monitoring and observability gap opening up. Most platforms can tell you what happened in a conversation. Almost none can reliably tell you why performance has changed or flag when a model update has degraded a specific intent. This is an underreported risk and one that buyers need to understand and navigate carefully.

More about our research

Our research represents one of the most comprehensive assessments of this vendor market available today. It was conducted entirely independently of any vendor relationship, and covers 17 leading vendors selected from a longlist of 60, assessed across 26 criteria including platform capability, integration and innovation.

It draws on 15 years of accumulated experience in the market, and is grounded in interviews with leaders running these platforms in production across aviation, financial services, retail, sports and consumer goods.

70 - 85%

CONTAINMENT
for well-defined automations

20 - 40%

COST-TO-SERVE
reduction in production

15 M+

INTERACTIONS
automated annually by the one airline

Honest answers before big decisions

Six questions to work through before you go anywhere near a vendor shortlist

The organisations in our research that have achieved the strongest results with AI agents share a common characteristic: they answered six questions honestly before they chose a vendor. The ones still running pilots three years in, without a scaled deployment to show for it, typically did not.

The six questions give you an honest picture of where your organisation is today: how much internal capability you have, how ready your data estate is, how clear your strategic mandate is. That picture determines your realistic deployment path, what timelines to plan against and which commercial model makes sense.

Skip the questions and go straight to vendor selection, and you may only find the answers mid-deployment – at significantly greater cost.

1 Do you have a clear strategy?

Strategic ambiguity is the most common reason pilots stall. No vendor can create clarity for you and no deployment reaches production without it.

Here's what good looks like: a named executive owner, a business case tied to specific measurable outcomes and sponsorship that can withstand difficult conversations about benefits, timelines or scope.

In our research: A global retailer put its successful deployment down to a well-defined use case: returns processed across three channels, with clear before-and-after metrics tracked from day one.

2 Do you have the capability to build, operate and improve an AI agent?

Most organisations underestimate what's required – not to build, but to sustain. A live agent needs conversation designers, integration engineers, product owners and data analysts. The technical capability to build is rarely the gap. The operational model to run the agent is.

In our research: One sports and entertainment organisation found that the human-in-the-loop process to keep the model accurate was constant and resource-intensive, and the internal team had neither the capacity nor the expertise to manage it. The technical build worked. The post-go-live model did not.

3 Have you chosen partners whose model fits your capability?

The engagement model matters as much as the technology. A vendor who takes full delivery responsibility needs a client who can define what good looks like, provide clean data and integrations, and make decisions quickly. A vendor who hands over a platform and steps back needs a client with strong internal engineering and conversation design capability. Misalignment between client and vendor is one of the most common causes of disappointing deployments and it rarely surfaces clearly in a sales process.

In our research: A major retailer ran its proof of concept with an outcome-based provider, but ultimately selected its existing enterprise platform partner – drawn by pre-built integrations, shared commercial infrastructure and an engagement model that matched the scale and complexity of its internal environment.

4 Is your data and integration estate ready?

An AI agent is only as good as the data it can access and the systems it can reach. A unified customer view, a documented API estate and a knowledge base structured for machine consumption are the essentials any vendor needs to deliver. Integration typically takes three to five times longer than building the agent itself, and the root cause of this is almost always inadequate data foundations, not technology.

In our research: One global e-commerce organisation attributed fast, straightforward integration to exactly this readiness. The platform it chose mattered, but the data it brought to the platform mattered more.

5 Do you have governance and risk clarity?

This question has two parts. Internally: do you have clear ownership, defined accountability and a governance structure with the authority to decide what the agent is permitted to do? Externally: what compliance obligations apply and have you resolved them before deployment begins? In regulated industries, the regulatory question is a filter. Everything else comes after.

In our research: A global financial institution's proof of concept stalled because a selected vendor couldn't confirm precise data residency locations – a regulatory requirement across multiple jurisdictions.

6 Have you defined the customer experience you're designing for?

Defining the customer experience means being explicit about how customers move between the AI agent and a human, what the transition feels like and which segments need different treatment.

Seamless handoff is harder than it appears. A customer who explains their situation to an AI agent and then repeats it to a human agent has a worse experience than one who reaches a human directly. Handoff design is a product decision that needs to be made before the platform is configured.

Inclusive design matters too. AI journeys built without considering customers with accessibility needs, lower digital literacy, non-standard speech patterns or vulnerability markers create legal and reputational risk. Under the FCA's Consumer Duty, identifying and routing vulnerable customers appropriately is not a feature to add later.

CASE STUDY SPOTLIGHT

When the vendor isn't the issue

A fast-growing sports and entertainment organisation selected a specialist AI platform on the basis of speed to deployment, peer referrals and cost profile. The platform performed well. What it couldn't solve was a knowledge and data challenge the organisation hadn't fully confronted: real-time data across four-day live events, multi-language requirements across eight countries and a small internal technology team with limited capacity for ongoing maintenance.

The agent required constant training. The ROI case was built around replacing ten agents with five supported by AI, but this has remained unproven. The lesson is not that the wrong vendor was chosen. It's that the six questions weren't fully answered before the decision was made.

Move ahead with confidence

With the six questions answered, you're in a strong position to choose the right way forward.

Our experience has shown this is always a balanced decision, weighing up speed, cost, quality and internal ownership. We have typically seen clients choose one of four routes based on the outcome they want to achieve.

Here are the four main options:

When your priority is building internal capability and maintaining control

- Use your CCaaS platform as the orchestrator and bring in the LLM, speech and knowledge components underneath.
- The most flexible path and the highest long-term optionality.
- Requires strong internal AI and engineering skills, or an established systems integrator relationship.
- Suits organisations with mature technology functions, significant contact volumes and a strategic intent to treat AI agents as a core capability.
- Realistic timeline: 6–12 months for a first production use case, assuming data and integration foundations are in place.

When your priority is delivering benefits quickly

- Use your CCaaS platform for telephony and routing, and outsource the conversational AI to a specialist provider with pre-built sector accelerators.
- Greater pace, lower up-front design burden, stronger alignment between commercial terms and delivered outcomes.
- Suits organisations with limited internal AI expertise, pressure to deliver visible results.
- Frame this as a three-year decision rather than a permanent one. Contract terms should protect the option to migrate to a different specialist or fold the capability into the CCaaS orchestrator as the platform layer matures.
- Realistic timeline: 8–12 weeks for a first scoped deployment using pre-built sector accelerators, assuming integration foundations are in place.

3 When you want to build foundations

- Use a fully managed or outcome-based provider – treat the first year as investment in foundations, not deployment.
- For organisations without a modern contact centre platform, a mature API estate or a knowledge base structured for AI consumption.
- Start with one well-scoped use case where volume is high, the process is straightforward and regulatory stakes are low.
- Agent co-pilot is a good start: it delivers measurable value, carries lower compliance risk than customer-facing automation, and builds the foundations you need for everything that follows.
- Realistic timeline: 12 months to get foundations in order and first use case into production.

4 When you're already running AI agents at scale

- Go further than customer service. Start connecting the front door to the back office: billing, field service, claims and payments handled end-to-end without handoff.
- Act now to ensure your services are visible to customers' AI agents. Adoption of MCP, which provides a common language for all AI models, is accelerating. You need clean, authenticated APIs to keep pace.
- Build differentiation by using interaction data to anticipate needs, personalise at scale and drive retention.
- Realistic timeline: the gap between prototyping and production is measured in months, not years.

A market full of promises. Here's how to read it

How to navigate a complex vendor market and find your fit.

Knowing whether your organisation is ready to deploy is the first question. Knowing which part of the market to engage with is the second – and it's harder than it sounds. Most analyses segment the market according to what vendors sell. We think it's more helpful to look at what role different types of vendor play for you and what you need to bring to make the relationship work.

Here are the four segments we see when we ask: 'who does what?' and 'what does the buyer need to be capable of to get value?'.

The Orchestrators are hyperscalers, CCaaS platforms and CRM-native providers that have positioned themselves as the orchestration layer of enterprise AI. Examples include AWS, Google, Microsoft, NICE, Cognigy, Genesys, Five9 and Salesforce.

The Orchestrators bundle LLM and orchestration into a managed service, offer bring-your-own speech components and absorb model-update risk as part of the platform. If you have an existing strategic relationship with one of these providers, this is often the path of least resistance. The trade-off is breadth over depth: conversational AI is one capability among many, not the core product.

Specialist AI platforms are independent, full-stack conversational AI vendors that plug into your existing CCaaS or telephony infrastructure as the conversational layer. Examples include PolyAI, Parloa, Druid AI, Replicant, Decagon and Fin.

This segment competes on platform quality, voice naturalness, agent-builder tooling, observability and the depth of enterprise integrations. Choose these it you want to retain your existing telephony backbone while adding a purpose-built AI layer on top of it.

Outcome-based challengers offer a full-stack conversational AI capability and charge on results rather than inputs. They exist in two forms: vertical specialists such as Gradient Labs, which accelerate time to value through pre-built sector use cases with tested guardrails; and generalists such as Sierra, which promise outcomes across any sector. A specialist's pre-built accelerators are genuinely valuable. A generalist's terms require you to contribute more design and optimisation work than the commercial structure implies.

Deep technical specialists lead with best-in-class capability at a single layer of the stack: speech-to-text or text-to-speech, where the gap between them and the rest of the market is measurable. Examples include Deepgram (STT) and ElevenLabs (TTS). Most platforms in the other three segments use these providers under the hood. However, you can engage with them directly – particularly when assembling a bespoke stack or when voice quality is a primary differentiator in your customer experience.

Assessing the vendors: findings from our 2026 research

As part of our 2026 research, we carried out an in-depth assessment of 17 leading vendors across six dimensions:

- Platform capability
  Agent builder, simulation tooling, observability, orchestration quality, LLM and speech optionality
- Integration & flexibility
  API surface, CCaaS integration, CRM integration, omnichannel coverage, custom development support
- Deployment & maintenance
  Sector agents out of the box, time to first deployment, continuous improvement, regression testing
- Trust
  Data security, PII handling, guardrails, vulnerability management, compliance posture
- Innovation
  R&D investment, patent activity, product roadmap, pace of new capability release
- Commercial & adoption.
  Pricing predictability, commercial flexibility, market traction, customer references

We tested every finding against interviews with leaders running these platforms in the real world, across aviation, financial services, retail, sports and consumer goods.

Here's what we found:

Trust is the dimension with the most significant and consequential gaps.

Vendors scored an average 3.6 out of five on trust – but this average is elevated by the orchestrators, whose infrastructure-grade compliance and data sovereignty commitments reflect years of enterprise procurement scrutiny. Remove them and the picture changes materially: several specialist and outcome-based providers score significantly lower, with data security and vulnerability handling the weakest areas. In regulated industries, trust is the filter to apply before you consider any other dimension. Outside them, it is a risk that rarely gets factored into vendor selection – but it should be.

Innovation separates the hyperscalers from everyone else.

The hyperscalers lead on innovation by a significant margin, reflecting genuine R&D scale. The deep technical specialists also score strongly, reflecting focused research investment in their core domains. Specialist AI platforms and outcome-based challengers score materially lower – which matters less than it might appear, since specialists are consuming and deploying hyperscalers' innovations rather than competing with them. What the gap does signal is that specialist vendors are more exposed to being leapfrogged at the component level. Ask what model flexibility looks like in practice, not in sales materials.

Deployment scores reveal the real trade-off between control and speed.

Outcome-based challengers score highest on deployment, because taking on delivery responsibility is their core proposition. Vertical specialists score particularly strongly: pre-built sector accelerators deliver faster time to value than assembling components from scratch. Orchestrators and specialist platforms cluster in the middle, reflecting that, in both cases, the buyer carries more of the design and integration burden. At least one CCaaS provider scores materially below the segment average, a flag for any organisation evaluating it as a path to fast value.

Integration scores favour the orchestrators.

Hyperscalers and CCaaS platforms average 3.7 against specialists' 3.4 and the outcome-based challengers' 2.9. Organisations standardised on a hyperscaler or CCaaS platform benefit from pre-built connectors that simply don't exist for independent specialists. If you don't have an existing strategic platform relationship, the integration gap is real and you will need to factor it into timeline planning.

Commercial scores are lowest for the outcome-based challengers

Despite the commercial model being the primary differentiator for outcome-based challengers, they score only 2.7 on commercial factors. This is because the commercial score captures whether pricing is predictable, commercial terms are clear and references are verifiable – not whether the model is attractive in principle. For example, one vendor's complex outcome definition, with negotiated 48-hour return rules and no standard pricing, scores poorly on predictability regardless of the model's theoretical appeal. Make sure to stress-test the commercial terms as rigorously as the technology.

For full details of how individual vendors scored, get in touch with one of our experts.

SECTION 04

Assessing the vendors: what we looked at

Platform capability	Agent builder, simulation tooling, observability, orchestration quality, LLM and speech optionality
Integration & flexibility	API surface, CCaaS integration, CRM integration, omnichannel coverage, custom development support
Deployment & maintenance	Sector agents out of the box, time to first deployment, continuous improvement, regression testing
Trust	Data security, PII handling, guardrails, vulnerability management, compliance posture
Innovation	R&D investment, patent activity, product roadmap, pace of new capability release
Commercial & adoption	Pricing predictability, commercial flexibility, market traction, customer references

Every vendor is different. The building blocks aren't.

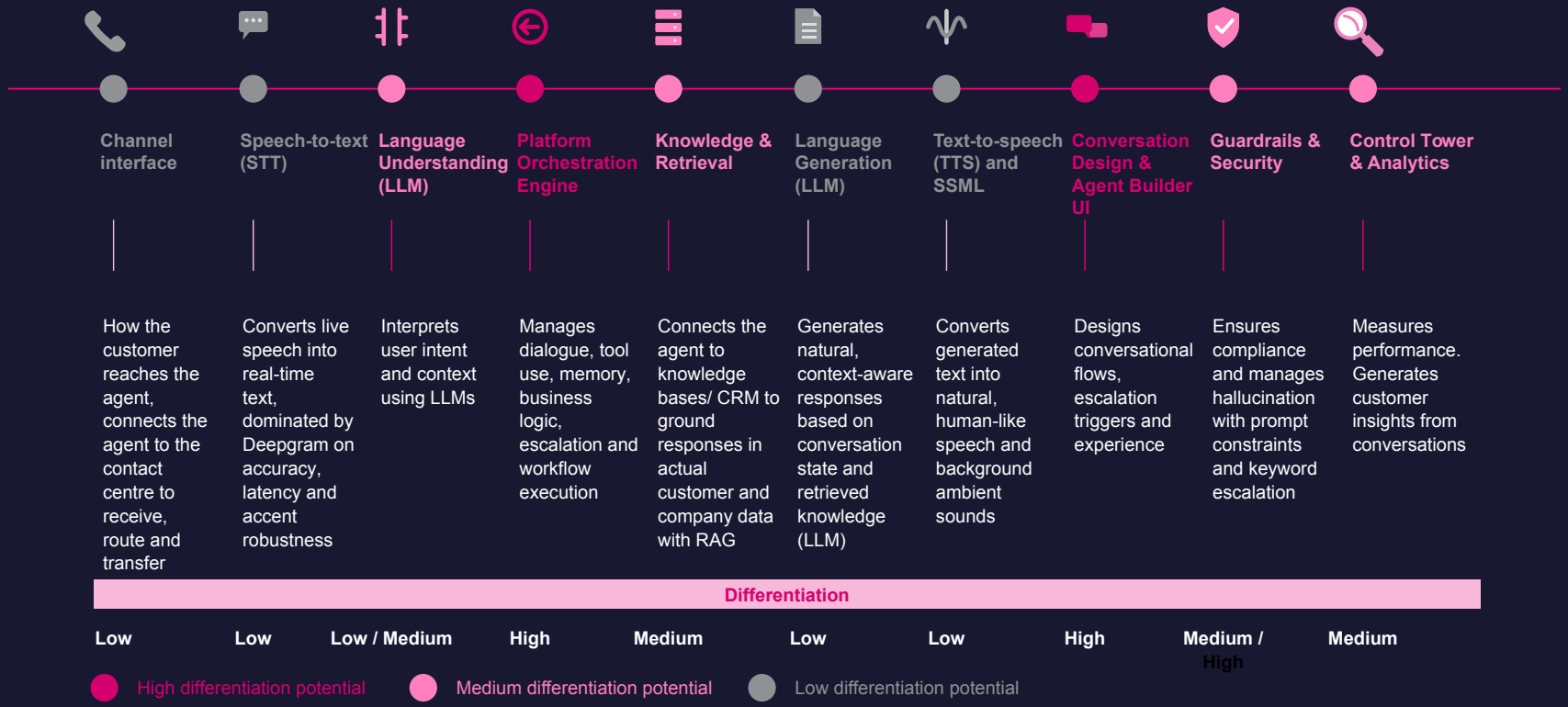
Understand what every AI agent is made of – so you can ask the right questions

Choosing the right vendor isn't about asking which one has the best AI. It is about understanding the architecture well enough to make a deliberate decision about which layers you control, which you outsource and what the consequences are when something needs to change. Getting this right starts with understanding what you're actually buying.

Switch the lens to see more clearly

Every AI agent solution is built on the same ten-layer stack (see page 18). What differs across vendors is who assembles those layers, who operates them, who bears the risk when something breaks and which layers are abstracted away from the buyer entirely. Those choices determine something that matters more than a feature comparison: how much you need to bring to deliver successful agents. **Every customer-facing AI agent is built on the same ten layers.**

Ten layers. Every agent. No exceptions



Layer 1 - Channel interface:
How the customer reaches the agent: voice, web chat, WhatsApp, mobile app, WeChat, email. The channel interface connects the customer to the contact centre infrastructure to receive, route and transfer.

Low differentiation potential: the channels themselves are broadly standard across the market.

Layer 2 - Speech-to-text (STT):
Converts live speech into real-time text for the agent to process. Deepgram dominates on accuracy, latency and accent robustness across approximately 50 languages and powers most platforms under the hood.

Low differentiation potential for the buyer: the layer is largely commoditised and can be swapped.

Layer 3 - Language understanding (LLM):
Interprets user intent and context using large language models. Gemini currently leads for customer service voice tasks where low latency is critical. Claude leads on reasoning. Mistral and Llama are strong options for private workloads.

Low to medium differentiation: the models are broadly available to every vendor, though platform flexibility to swap them varies.

Layer 4 - Platform orchestration engine:
Manages dialogue, tool use, memory, business logic, escalation and workflow execution. This is the brain of the agent: it tracks where the conversation is, decides what to do next, calls back-end systems and determines when a human needs to step in.

High differentiation potential: this is the layer where vendor choice creates the most durable advantage and where the gap between a convincing demo and a reliable production deployment is most exposed.

Layer 5 - Knowledge and retrieval:
Connects the agent to knowledge bases and CRM to ground responses in actual customer and company data, typically via retrieval-augmented generation (RAG).

Medium differentiation: the retrieval technology is broadly similar across leading platforms. What differs is the quality of the knowledge base the organisation provides. Poorly structured content written for human agents does not translate to AI retrieval, and this is the most common cause of containment failure.

Layer 6 - Language generation (LLM):
Generates natural, context-aware responses based on conversation state and retrieved knowledge.

Low differentiation: this layer uses the same frontier models available to every vendor. The quality of the output is determined more by the orchestration and knowledge layers than by model choice at this point in the stack.

Layer 7 - Text-to-speech (TTS) and SSML:
Converts generated text into natural, human-like speech, with SSML providing control over pacing, emphasis and ambient sound. ElevenLabs leads on voice naturalness and expressiveness and is the de facto standard across much of the market.

Low differentiation for the buyer: the layer can be swapped and most platforms allow bring-your-own TTS.

Layer 8 - Conversation design and agent builder UI:
The environment in which conversations are designed, tested, maintained and improved.

High differentiation potential: on some platforms every change requires a developer; on others a business analyst or contact centre manager can make updates directly. This difference compounds significantly over time in cost, speed and the organisation's ability to improve the agent continuously after go-live. It is the most underweighted criterion in most vendor RFPs.

Layer 9 - Guardrails and security:
Ensures compliance and manages hallucination risk through prompt constraints, keyword escalation triggers and data protection controls.

Medium to high differentiation: this is where the widest gaps exist between vendors. Several providers with strong platform capability score poorly on data security, PII handling and vulnerability management. In regulated industries this layer functions as a filter before any other evaluation begins.

Layer 10 - Control tower and analytics:
Measures performance, generates customer insights from conversations and supports continuous improvement.

Medium differentiation: the tooling exists but is immature across most of the market. Most platforms currently offer dashboards that track what happened. Few reliably surface why performance has changed or flag when a model update has silently degraded a specific intent.

Where the real difference lives

Some layers of the stack are broadly the same whoever you buy from. A few aren't – and those are the ones worth paying attention to.

Orchestration is the first layer of genuine advantage. The orchestration engine tracks the conversation, understands what the customer is trying to achieve, connects to the back-end systems and decides when to escalate to a human.

A weak orchestration layer produces agents that feel capable in a demo and unreliable in production. The pattern that has produced the strongest results in our research is hybrid architecture: deterministic workflows for safety-critical or high-stakes processes, LLM-powered orchestration for everything else.

Agent builder and conversation design tooling is the second layer of lasting differentiation – and the most underweighted criterion in most vendor RFPs. The quality of the tools used to design conversations, simulate interactions and improve the agent post-launch determines long-term outcomes and total cost of ownership more than any other factor.

Once an agent is live, someone needs to maintain it, extend it and improve it as the business evolves. The question is who that person needs to be. On some platforms every change requires a developer. On others a business analyst or contact centre manager can make updates directly.

That difference compounds over time in cost and speed.

SECTION 05

The layer no-one has cracked yet

Monitoring and observability is the layer the market hasn't solved yet, and it's the one with the most direct impact on whether deployments sustain their results over time.

The core problem is this. Deterministic systems are easy to monitor because you know what should happen and can check whether it did. Probabilistic systems, which is what every LLM-powered agent is, don't have a fixed sequence of steps to validate against. A global airline described the challenge precisely: checking every few minutes whether a probabilistic system is still behaving as designed requires a fundamentally different approach from the automated regression testing that works for deterministic flows.

Most platforms currently offer analytics dashboards that track containment, escalation rates and conversation volume. These are useful but not enough – they tell you what happened, not why, and won't flag when a model update has silently degraded performance on a specific intent.

Treat observability requirements as a contractual requirement, not an afterthought: what data can be extracted, in what format, at what granularity, and what SLAs apply to incident detection? Design for external monitoring from the start. And assume that the first sign of a performance problem will be a shift in containment rates or escalation volume, not a vendor alert.

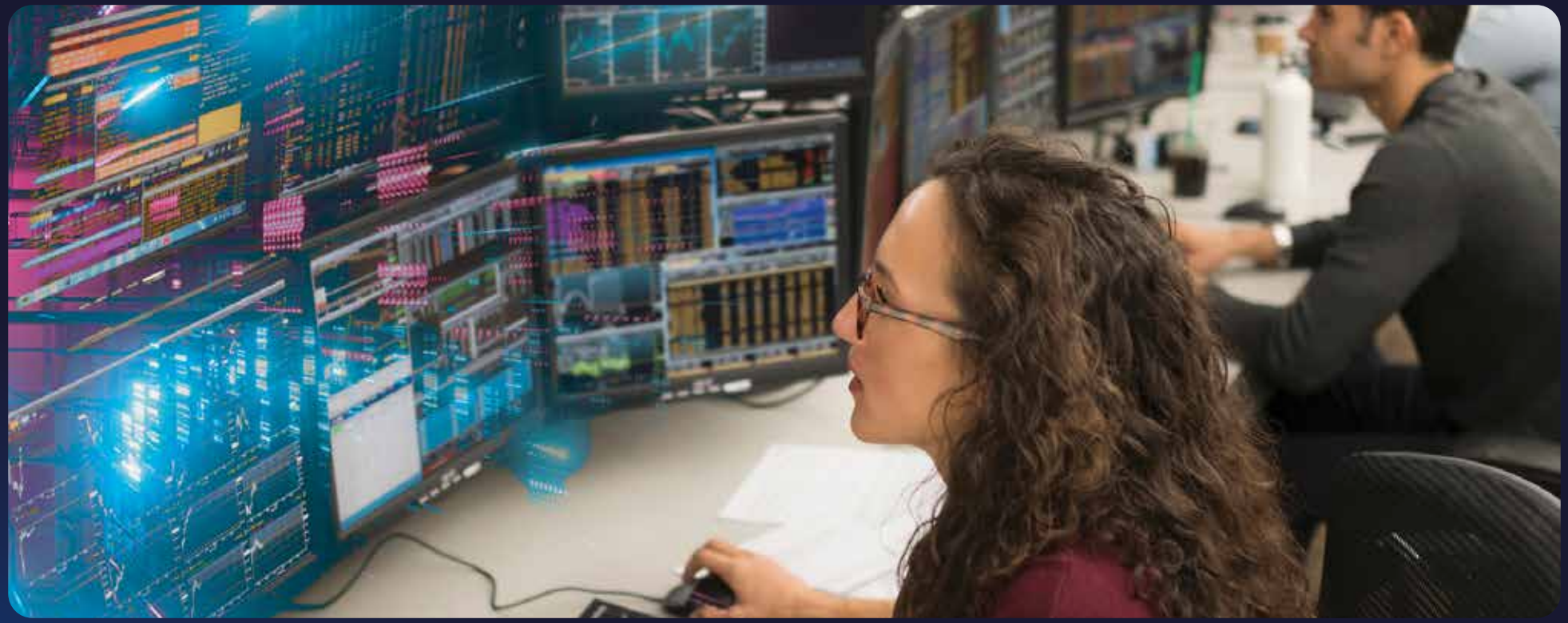
The demand for independent third-party conversation monitoring is real and vendors who solve it will gain meaningful advantage. For now, it is the layer where the market is most exposed and buyers are least protected.

To find out which vendors create lasting differentiation, talk to one of our experts.

How hands-on do you want to be?

The level of control you have over each layer depends on which market segment you engage with.

- **Orchestrators** offer the most flexibility, leaving you to assemble components and manage the integration burden
- **Specialist platforms** bundle more of the stack but allow you to keep configurability
- **Outcome-based challengers** abstract most of the complexity, taking on delivery responsibility but leaving you with less control
- **Deep technical specialists** power individual layers within other vendors' stacks and are rarely engaged directly.

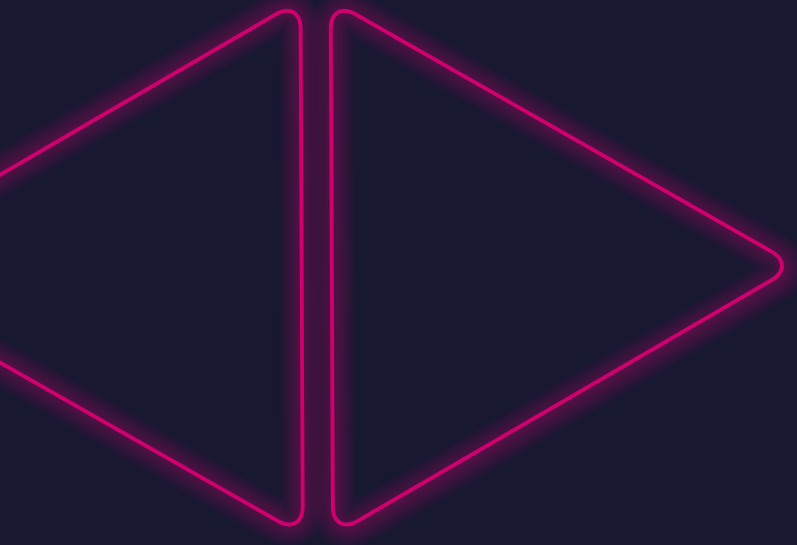


Getting the deal right

Usage-based, platform or outcome-based – which model fits your reality?

Every commercial model comes down to two things: how you pay, and how much you build and operate yourself. The two are inseparable. A low per-minute price is meaningless without the engineering team to turn components into a working agent.

A high per-outcome price may represent better value if the vendor takes on the delivery responsibility that you'd otherwise carry internally.



The headline price is just the beginning

Whichever commercial model you choose, the headline price captures only a fraction of the true cost.
The largest cost driver in every model is the people you need to design, operate and improve the agent.

- **Conversation design and user research:** underinvestment here is the most common reason containment targets get missed. Getting the conversation architecture right before build isn't optional – organisations that have treated it as such have consistently found out after go-live, at greater cost than doing it right the first time.
- **Integration and development:** plan for this to take three to five times longer than building the agent itself. Here organisations have invested in a unified Customer Action Gateway - a stable interface that abstracts back-end complexity - they recover the cost quickly in reduced engineering time and lower regression risk.
- **Regression testing and model management:** is a persistent operational cost that most organisations underestimate at procurement. When LLMs update, conversation flows can break without warning.
A global airline in our research described even minor model version changes causing tool-calling failures from the first turn. Automated regression testing across the full conversation surface is a core operational dependency.
- **Quality assurance and monitoring:** requires continuous assessment of conversation quality, accuracy, sentiment and containment.
The observability tooling to do this independently of the vendor is still immature – but build the requirement into contracts now.
- **Compliance and governance:** costs are highest in financial services, where PII masking, data residency rules and full auditability of every tool call may all be mandatory. LLMs may not be permitted. These are not marginal line items.
- **Change management:** redefining contact centre agent roles, training on escalation handling and the cultural shift from handling to supervising is consistently underbudgeted and consistently consequential when it is. It belongs in the total cost model from the start.

SECTION 06

Three ways to pay

Usage-based models: lowest unit cost, highest engineering bill.

How you pay

Deep technical specialists and hyperscalers typically charge on consumption – per minute of audio processed, API call or token generated.

Pricing varies by component:

- Speech providers typically charge fractions of a penny per minute.
- Conversational AI platforms with bundled orchestration charge roughly 8p –13p per minute.
- Language model inference can add 10–30 per cent, depending on model and conversation complexity.

Total spend is harder to predict than it appears. The per-minute price covers only the compute.

How much you do yourself

- You assemble your own stack, selecting best-of-breed components and integrating them.
- Conversation design, integration engineering, regression testing, ongoing optimisation and operational support all sit with you.
- For a usage-based deployment, you'll need, at a minimum, conversation designers, integration engineers, a data analyst and a product owner, plus cloud engineers with platform-specific expertise or a systems integrator relationship.

Selected vendor highlights

- AWS is a good example of an end-to-end build option when you want to control the full stack.
- Bedrock is secure, data is ringfenced within the customer account and workloads can be kept within specific regions without complexity.
- Google Cloud's CCAI and Gemini are compelling where voice is the priority.

SaaS platform models: higher unit cost, lower engineering bill

SaaS platforms bundle AI agent capability within broader CCaaS, CRM or enterprise software platforms, absorbing a significant share of the engineering burden as a result.

How you pay

Pricing is per conversation or per agent seat, at rates higher than usage-based models. The premium pays for infrastructure and tooling you don't need to build yourself. On the better platforms, the underlying language, speech and voice components remain configurable.

How much you do yourself

A SaaS deployment needs a smaller, more design-focused team: conversation designers, a product owner, a business analyst and a platform administrator, with integration developers brought in only for bespoke back-end systems.

Selected vendor highlights

On the better platforms, the underlying language, speech and voice components remain configurable. Genesys is a clear example: the Large Action Model runs on underlying AWS infrastructure. AWS is a good example of an end-to-end build option when you want to control the full stack.

Bedrock is secure, data is ringfenced within the customer account and workloads can be kept within specific regions without complexity.

Where user adoption through a voice channel is the priority it is hard to find examples which are Orchestrator platforms only

Outcome-based models: pay for results, not input

How you pay

Outcome-based models charge on results: a resolved enquiry, a completed transaction, a contained interaction. The commercial logic is appealing – incentives are aligned and the vendor only earns when the customer succeeds. But the reality is more nuanced and the definitions are worth stress-testing before you sign.

The first complication is agreeing on what counts as an outcome. Definitions vary significantly across vendors and are negotiated individually rather than set by any standard. For example: if a customer returns within 24 to 48 hours with the same enquiry, the vendor may not charge again – but a return after 48 hours with the same question counts as a new outcome not a failed one.

The commercial model is less structurally different from usage-based or SaaS than vendor marketing suggests. Model the full cost under realistic volume assumptions and compare like for like.

How much you do yourself:

With an outcome-based provider, the delivery responsibility genuinely shifts: the vendor only earns when the customer succeeds, so design, optimisation and ongoing improvement are in their interest to own. The caveat is for generalist providers, these require significantly more client involvement than the commercial structure implies to bring the sector domain knowledge, making the internal capability requirement closer to a SaaS deployment than the pricing model suggests.

How different vendors compare

Specialist providers such as Gradient Labs bring pre-built sector use cases with defined parameters for deterministic versus reasoning behaviour, established guardrails and tested escalation paths. Time to value is materially faster and the delivery responsibility genuinely shifts to the provider. Generalist outcome-based providers offer the same commercial structure without the same sector depth. Journey design, conversation architecture, knowledge curation and ongoing optimisation require you to be much more involved, making the practical internal capability requirement closer to a SaaS deployment than the pricing model implies. The outcome-based label describes the commercial structure, not the effort required.

In short: how different models compare

	Usage-based	SaaS platform	Outcome (specialist)	Outcome (generalist)
Pricing unit	Per minute	Per conversation	Per resolved outcome	Per resolved outcome
Indicative voice cost	Lower per interaction	Higher per interaction	Premium per outcome	Premium per outcome
Engineering needed	High	Moderate	Moderate (domain-accelerated)	Moderate
CX design	Fully client owned	Client owned, platform tooled	Provider-led	Shared
Speed to deploy	Months	Weeks to months	Weeks	Weeks to months
Vendor lock-in	Low	Moderate to high	High	Moderate

What does ready even look like?

And how to work out if you're there.

Earlier, we introduced six questions to help assess how ready you are to deploy an AI agent. Now let's look at what good looks like across each of the four building blocks of AI readiness – and what to do if you're not there yet.

Strategic clarity: Strategy and go-to-market

In our research, organisations that achieved the strongest results treated deploying an AI agent as a business transformation that happens to involve technology, rather than a technology project looking for a business case. That might sound like a subtle difference, but it has a big impact on how your programme is structured, who owns it and how decisions get made when things don't go to plan.

To be ready, you need: commercial goals tied to measurable outcomes, a prioritised set of use cases and sponsorship by someone senior enough to make resourcing decisions and keep momentum behind them. It's common to see AI activity springing up across different business units at the same time, each with its own vendor conversations and pilot projects. But that's also where many programmes get stuck and never make it to production scale.

If strategic clarity is weak, the priority before talking to vendors is to resolve two things: what are you are trying to achieve, expressed in terms that can actually be measured, and who is accountable if it doesn't happen.

Operational capability: Structure, skills, culture and change

There are three dimensions to think about here: how your business is structured and what capabilities it has, whether your organisation has the right skills and talent, and whether your culture is ready for change. Together they determine whether you can build, operate and continuously improve an AI agent after go-live. That ongoing capability is one of the biggest factors separating successful deployments from proof-of-concepts that never make it into production.

Structure The question isn't just whether the right roles exist. It's whether those roles are set up to work effectively together. Conversation designers disconnected from contact centre operations, integration engineers sitting outside the AI programme or product owners with the mandate but not the budget – these misalignments are as damaging as capability gaps. Organisations running AI agents successfully tend to build dedicated cross-functional teams with clear accountability for outcomes, not just for delivery.

Skills If you're deploying through a specialist platform or outcome-based provider, you don't need a large, in-house AI engineering team. What you do need is a level of AI literacy among the people who own and run the agent. That means people who can assess conversation quality, spot recurring failure patterns, brief improvements accurately and hold vendor partners to account. The good news is that these skills can be learned. Organisations that invest in them early typically recover the cost through faster iteration and fewer issues after launch.

Culture and change This is where many technically sound deployments struggle to gain traction. Redefining agent roles, being transparent about what AI will and won't change, and helping teams move from handling interactions to supervising AI systems all require sustained investment.

In our research, organisations succeeding with AI treated frontline teams as co-creators of the AI, not just end users. When agents understand why the AI makes the decisions it does, contribute to improving it and see it as something that supports rather than replaces their role, the results are consistently better. These teams maintain higher quality, identify issues earlier and contribute more effectively to improving containment than teams who weren't involved.

Data and integration: Data, systems and process

There are three key readiness areas here: data and insights, systems and tools, and process and policies. Together, they are some of the strongest predictors of whether an AI deployment succeeds in the first 12 months. They are also where the gap between perceived readiness and actual readiness is often the largest.

Data The standard you need to reach before starting is a unified view of the customer that can be accessed through stable, documented APIs, with clean identity and authentication architecture that works for machine-to-machine interactions. The reality is that most organisations aren't there yet.

The practical response is not to wait for perfect conditions. Instead, start with use cases that can work with the data you already have in usable form. Agent co-pilot, which brings up valuable knowledge for a human agent during a live call, is a good first step because it draws on a single knowledge source, doesn't need deep integration into back-end transaction systems and generates the containment data that informs everything that follows.

Knowledge base quality This was the single most underestimated factor across every organisation in our research. Content written for human agents often includes informal language and implicit escalation cues that don't translate well for AI. The standard to aim for is simple: every knowledge article should clearly define the situation, the action required and the expected outcome, without relying on assumed context.

Restructuring a knowledge base for AI consumption is time-consuming, but it's also one of the highest-return investments you can make before deployment. Organisations that do this work report significantly better containment rates and fewer post-launch issues than organisations that simply point the platform at their existing content and hope for the best.

Systems integration This is another area where expectations are often unrealistic. In many cases, integration takes three to five times longer than building the agent itself. The reasons are consistent: undocumented API dependencies discovered mid-project, inconsistent interfaces across back-end systems and authentication architectures that were never designed for machine-to-machine interaction.

The most effective way to speed things up is with a unified Customer Action Gateway: a stable integration layer that abstracts back-end complexity and gives the AI agent a single clean interface. Organisations that invest in this approach typically recover the cost quickly through reduced engineering effort and lower regression risk when back-end systems change.

Process AI cannot act on knowledge that exists only in people's heads. If a process relies on an experienced agent using judgement or context to know when to escalate, that context needs to be made explicit before any AI agent can be trained on it. Documented, standardised workflows – where rules are written down rather than passed down through experience – are essential for anything beyond basic FAQ automation. Interestingly, organisations that do this work often report an additional benefit: the process improvements required for AI deployment frequently improve human agent performance as well.

Governance and risk: Governance and compliance

This dimension has two parts: internal governance and regulatory compliance. You need to resolve both before you start deployment, not work them out along the way.

Internal governance. The fundamentals are straightforward: clear ownership, clear accountability and a cross-functional forum with the authority to make and enforce decisions about what the agent can do, how it is monitored and when escalation is required.

Without these foundations, risk accumulates quietly. Edge cases are not addressed, model updates are not properly regression-tested and the agent gradually drifts away from its intended behaviour without anyone noticing. In many cases, the problem only becomes visible when a customer or a regulator spots it.

Governance gaps expand as deployments grow. This means that if you treat governance as a one-time procurement check rather than an ongoing operational discipline, you likely to accumulate risks that remain invisible until they become serious problems.

Regulatory compliance. This requires the same serious attention as internal governance. The obligations that matter most for UK organisations are:

- UK GDPR – this applies to AI-transcribed and AI-processed voice interactions because the data involved is personal data. You must establish lawful basis for processing, ensure data minimisation principles apply across all vendor systems through which data flows, and have clear policies covering retention and deletion.
- FCA obligations apply to AI-handled interactions in financial services in exactly the same way they apply to human-handled ones. Consumer Duty requirements, vulnerable customer identification and recording obligations all extend to AI agents.
- The EU AI Act applies to UK organisations serving EU customers. Customer-facing AI agents in sectors such as financial services, insurance and healthcare are likely to fall into high-risk categories, bringing additional requirements around conformity assessment and human oversight before deployment.

Transparency is increasingly moving from best practice to regulatory expectation. Several vendors now provide out-of-the-box disclosure flows, and in some sectors informing customers that they are interacting with AI is already mandatory.

One lesson came through repeatedly in our research: resolve data residency before selecting a vendor selection, not after. A vendor that can't give a clear and precise answer on where customer data will be processed and stored cannot clear procurement in a regulated environment, regardless of how strong the product is. The cost of discovering this too late is significant. Several organisations in our research described proof-of-concept programmes that stalled for more than a year because data residency concerns emerged after vendor selection rather than before it.

INDUSTRY SPOTLIGHT

Readiness as a strategic investment

A global commercial real estate firm assessed its readiness across all dimensions before launching AI deployment in its investor accounting and reporting function. The review found strong leadership commitment and deep process expertise, but also highlighted fragmented data, limited enterprise governance, processes that relied heavily on organisational memory rather than documented rules, and uneven AI literacy across regions.

Rather than jumping straight into vendor selection, the organisation spent its first six months focused on three priorities: assessing data readiness for priority use cases, establishing a cross-functional governance forum, and building alignment through targeted communications.

The result was a clearer roadmap for the deployment, including the foundational investments needed to accelerate the programme in year two.

The shape of things to come

From conversational interfaces to customer-owned AI, the next stage is already taking shape.

The technology decisions you make today are not just about improving customer service. They're about establishing the architecture for a fundamentally different mode of customer interaction. Get this right and you won't just have better contact centres. You'll have the infrastructure for a new commercial relationship with your customers

The intelligent front door: one experience, everywhere

Customer interaction is becoming conversational across every channel. Search boxes, help menus, product configurators, onboarding flows, app navigation and in-product support are all moving from forms and menus towards natural language.

At the same time, most organisations are still running separate bots for each channel: a voice bot, a web chat, a WhatsApp agent. But customers don't think in channels. They think in outcomes. They want to ask a question and get it resolved, regardless of where they ask it.

This is where the intelligent front door comes in – a single, channel-agnostic conversational layer that sits behind every channel, powered by a one orchestration engine, one knowledge base and one set of guardrails. When the knowledge base is updated or a guardrail is changed, every channel reflects it.

If you're planning your second or third AI deployment, this is the architecture you should be designing for.

Customer owned AI: the bot on the other side

A parallel shift is underway that will reshape customer service even more profoundly: the emergence of customer-owned AI agents acting on behalf of individuals.

The Model Context Protocol (MCP) provides the open standard that makes this possible. It allows an AI agent to connect directly to external services, authenticate on behalf of a user and complete transactions autonomously. If your services are locked behind human-only interfaces, they will be invisible to AI-mediated interactions.

ChatGPT already supports direct integration with commercial services through this standard, reaching over 800 million active users every week. A global airline in our research is already prototyping MCP-based integrations. And a major investment bank has built MCP server infrastructure internally and will extend it to external AI agents once the authentication architecture is ready to go.

Overall, our research suggests meaningful customer-AI interaction volumes within 12–24 months in retail and travel, 24–36 months in financial services and 36–48 months in government.

What does this mean in practice? Avoid building separate MCP infrastructure. The same Customer Action Gateway that supports your customer-facing AI agents is the foundation for customer-owned AI.

Build your tool servers as MCP endpoints and both internal and customer-owned agents can use them.

> **“Build the orchestrator, capability agents and tool servers once, and the same architecture serves every channel – including the customer’s own AI.”**

Market consolidation: fewer vendors, bigger platforms

The vendor landscape will look very different in the next 24 months.

Orchestrator-specialist integration will accelerate. NICE's acquisition of Cognigy and AWS's acquisition of NLX are early examples. As the market matures, orchestrators are looking to strengthen their AI capabilities, while specialists need greater reach into enterprise buyers. If you're considering a specialist platform, assess how likely it is to be acquired and what that could mean for commercial terms and product roadmap.

Industry specialisations will also become more important. Outcome-based generalists face pressure from both sides: vertical specialists with deep industry expertise and pre-built accelerators that deliver faster time to value, and orchestrators with broader integration capabilities. When evaluating vendors, make sure sector depth is genuinely built into the offering rather than something you need to provide yourself.

Many vendors in this market are venture-backed businesses less than five years old, running on growth capital with no clear path to profitability. Enterprise buyers are concentrating spend on providers with credible production track records. The cost of choosing a vendor that is acquired, repositioned or fails before your deployment is complete, is significant. A vendor's financial health and long-term viability should be part of your evaluation process.

Build for the organisation you're becoming

- **Make deliberate decisions about architecture, not just technology**
- **Choose platforms that give control over the layers that matter**
- **Build integration foundations that can support multiple use cases**
- **Think about your second and third deployments before your first one goes live**

Need a partner to help you get it right?

We help you move from AI experimentation to production-scale value with confidence

The technology is no longer the constraint. What matters now is how well you design for it: the quality of your use case selection, the strength of your data and integration foundations, the discipline of your vendor choices, and the operating model that supports the agent after go-live.

These are the decisions that determine whether AI becomes a scalable business capability or remains a promising pilot. They're also the decisions organisations often find hardest to make. That's where we can help.

1 AI capability and readiness audit

We assess your readiness for an agentic architecture, from data and APIs to knowledge management and operating model design. You'll get a clear picture of where you are today, where the gaps are and what to do next.

2 Vendor selection and RFP support

We help you navigate a complex vendor market, from shortlisting providers and designing RFPs to evaluating commercial models and negotiating terms. You'll get an independent view of your options backed by our in-depth assessment of 17 leading vendors, the trade-offs involved and the partner best suited to your needs.

3 Use case identification and proof of concept delivery

We help you identify the highest-value AI opportunities, validate them through focused proofs of concept and build a credible path to production. You'll get a prioritised roadmap, the right technical foundations and a clear plan for scale.

4 Delivery acceleration with deep sector expertise

We combine sector expertise with proven delivery experience to help you avoid common pitfalls and accelerate time to value. You'll get access to tested architectures, established operating models and industry-specific insights that support long-term scale.

Let’s talk

We’re passionate about AI in service, with a long track record of working with clients across sectors including financial services, energy and utilities, retail, telecommunications, aviation and the public sector. Let’s talk about what we could do together.

Key Contacts

Carrie Hegan, Partner and Customer Practice Lead
carrie.hegan@baringa.com

Daniel Pitcher, Director and AI in Service Lead
daniel.pitcher@baringa.com

Sam Atkins, Partner and AI-Enabled Operations Lead
sam.atkins@baringa.com