



SACHA OMON AI
WE SELL FOR YOU



SACHA
AI SDR

Agentic AI + RAG



WEB



WHATSAPP



PHONE

WHITE PAPER

How Agentic AI and RAG Are Entering the Enterprise Right Now

RAG — Retrieval-Augmented Generation

144
pages

WHY NOW? • HOW IT WORKS • HOW TO DEPLOY

Salomon Bellaïche
Founder, SACHA OMON AI

INTERNATIONAL EDITION 2026

Data as of August 19, 2026: sources fully referenced in the appendix.

www.sacha-omon.ai



SACHA OMON AI
WE SELL FOR YOU

INTERNATIONAL EDITION • 2026

White Paper

How AI powered by agentic AI and RAG is now entering the enterprise

RAG — Retrieval-Augmented Generation

WHY NOW?
HOW DOES IT WORK?
HOW TO DEPLOY

Salomon Bellaïche

Founder of SACHA OMON AI
Editor-in-Chief of the White Paper • Publisher

INTERNATIONAL EDITION 2026 • EDITION NO. 1

U.S. ENGLISH EDITION • REVIEW NO. 1 • SEPTEMBER 1, 2026

Data current as of August 19, 2026: sources fully referenced in the appendix.

www.sacha-omon.ai

Table of Contents

This White Paper explains how AI is now entering the enterprise through the combination of agentic AI and RAG - Retrieval-Augmented Generation - to understand, access enterprise knowledge, and act.

Need a definition? The glossary on page 114 explains the technical terms used in this White Paper in plain language. Sources and cited articles are collected on pages 118 to 129.

FOUNDER'S STRATEGIC OPENING

4

How AI, powered by agentic AI and RAG, is now entering the enterprise	4
Why this White Paper exists - and the augmented organization	7
Why now? Yesterday you had to be found; today you have to act	10
The great shift	11
Reading guide: who this White Paper is for	14
The three conditions for action	15

PART I · WHY NOW?

26

01. The enterprise AI paradox	26
02. The platforms shaping enterprise AI	29
03. Customer cases and quantified evidence	38

PART II · HOW IT WORKS

42

A. UNDERSTANDING THE BUILDING BLOCKS

04. Understanding LLMs: the AI you know - and its limits	42
05. Understanding agentic AI	48
06. Understanding RAG: Retrieval-Augmented Generation	55
07. Agentic AI + RAG + LLM: why the combination wins	62

B. FROM REQUEST TO COMMERCIAL ACTION

08. From demand to immediate sales action	70
· From analysis to application: the multichannel agentic AI SDR agent	76

A. FRAME YOUR PROJECT

The 7 structuring questions for your project requirements document	84
From your answers to a project requirements document with your own AI	85
Stair-step deployment and the 8 rungs	86
The 14 decisions for your AI agent	87
Have your own AI draft your project requirements document	88

B. DEPLOY ACROSS THE ENTERPRISE

09. How to build and deploy an AI agent in the enterprise	93
10. Governance, security, and compliance	99
11. Mistakes to avoid before and during deployment	103
12. Where value is lost - and how it comes back	104
Decision summary: what you need to retain and decide	105

ABOUT SACHA OMON AI

108

WHITE PAPER APPENDICES

113

Appendix A - Overview of model providers and their engines	113
Appendix B - Glossary: terms and definitions	114
Appendix C - Methodology	117
Appendix D - Clickable sources and articles	118
Appendix E - Clickable sources for the strategic opening	127
Appendix F - Milestones 2017-2026: how modern AI was built	130
Legal notice	142
After the White Paper	143

FOUNDER'S STRATEGIC OPENING

Why this White Paper exists, what is changing, and the three conditions for action.

How AI, powered by agentic AI and RAG, is now entering the enterprise

FOUNDER'S OPENING

SALOMON BELLAÏCHE

Founder · SACHA OMON AI

Editor-in-Chief · Publisher

Expert in digital transformation and digital acquisition

More than 30 years of experience · More than 3,000 companies supported

From small businesses to global enterprises · Microsoft · L'Oréal



Salomon Bellaïche, founder of SACHA OMON AI.

WHITE PAPER CREDITS

Thinking, concept, and editorial direction: Salomon Bellaïche.

Research, structure, and verification: assisted by Anthropic's Claude and OpenAI models.

The starting point: protect the value created by customer acquisition

In 2025, Google and Meta together reported \$490.9 billion in global advertising revenue in their official filings. In the first half of 2026, that total reached \$273.3 billion, up from \$226.2 billion a year earlier - an increase of 20.8% and more than \$1.5 billion per day. These figures show the scale of investment devoted to capturing attention and creating intent.

But for an advertiser, intent is not yet revenue. Value can be lost precisely after the click, call, or message. A prospect arrives on the website: the company must respond immediately, understand the request, provide a relevant first answer, qualify the need, and prepare the commercial handoff. If the company is slow to act, intent cools before the sales team can even step in.

This is where an agentic AI sales agent can intervene. Connected to enterprise knowledge and teams, it responds, engages, qualifies, handles initial objections, and hands off an actionable opportunity. AI does not replace sales: it protects the value created by acquisition, while humans retain judgment and decision-

making. This conviction is grounded in more than 30 years of experience in digital acquisition, lead generation, and ROI management: performance is measured through revenue, not just the click.

Direct sources: Alphabet, 2025 10-K and H1 2026 10-Q · Meta Platforms, FY 2025 results and H1 2026 10-Q.

1996: building acquisition systems, not just campaigns

My core entrepreneurial journey began in 1996, when I founded my first Internet startup, among the early French companies focused on digital acquisition and demand generation. Since then, I have designed acquisition, lead-generation, and ROI-optimization programs for more than 3,000 companies, from small businesses and SMEs to global enterprises including Microsoft and L'Oréal. That experience taught me to look at the entire chain: the promise, the media, the landing page, the request, the sales response, qualification, and conversion into revenue.

The question running through these three decades has not changed: how do you turn marketing investment into real demand, then turn that demand into a relationship and revenue? Advertising tools have become more powerful, precise, and fast. Yet the experience after acquisition is still often built around a form and a delayed callback. We have perfected demand creation without always perfecting how we receive and act on that demand.

A form records interest; it does not handle it

A form is useful for collecting information, but it understands neither intent, urgency, nor objection. It does not advise, and it does not answer the question that triggered the visit. The Salesforce example is striking. In August 2025, Marc Benioff, Salesforce co-founder, Chairman and CEO, said that more than 100 million prospects who had contacted the company during its first 26 years had never been called back because there was not enough human capacity. If the world's leading CRM vendor faces this break in the chain, no organization can treat the issue as secondary.

Direct source: Marc Benioff, The Logan Bartlett Show, Episode 149, August 29, 2025. The full case is detailed in Chapter 3.

Bringing AI into the enterprise

LLMs opened a decisive new chapter. Through APIs and access keys, their capabilities for understanding, reasoning, and language generation can be connected to enterprise systems. But individual use of ChatGPT or Claude, however impressive, is not yet enterprise integration. For an executive, the real question is different: how do you connect this intelligence to enterprise knowledge, rules, channels, CRM, tools, and teams without exposing information that must remain protected?

That observation led me to create SACHA OMON AI. I chose not to begin with a broad productivity narrative. I started where the loss is visible, measurable, and directly tied to revenue: digital acquisition. The goal is to connect the moment a request appears to a reliable, immediate, contextualized response, then hand the human team a situation that is already understood and documented. To build this approach, I spent more than three years immersed in different call centers in Israel, observing conversations, objections, transfers, and the precise moments when knowledge is missing.

This work is part of nearly four years devoted to SACHA OMON AI. More than thirty years of experience in digital, acquisition, development, and ROI management allowed me to identify the business problem before all of the technology building blocks were mature, then follow their convergence: LLMs to understand and formulate, agentic AI to pursue an objective, RAG to ground the system in enterprise

knowledge, and multichannel capabilities to deliver a continuous experience. The diagrams and summaries in this White Paper reflect that thinking and the design choices behind SACHA OMON AI; they explain the principles, not proprietary mechanisms.

Agentic AI + RAG: the agent acts with enterprise knowledge

We also need to be precise about a word that creates confusion: automation. Automation means chaining tasks according to rules written in advance: if A happens, do B, then C. The system can be highly sophisticated and connect multiple applications, but the path remains predetermined - it executes the process it was given; it does not reason. I recently saw someone present a simple re-automation of business processes with a workflow tool as "AI." It is not the same thing: automation is execution.

A chatbot adds a conversational interface to that same logic. Historically, it follows a predefined scenario and breaks down when the user steps outside it; with language models, it understands requests better and produces more relevant answers. But conversation is not action: a chatbot can explain what should be done without having the autonomy, tools, or permissions required to do it. This is the kind of tool that left many users with bad memories - to the point that, for much of the public, "chatbot" still suggests a closed loop and off-target answers.

An AI agent introduces a different logic: it receives a request, retrieves what it needs from enterprise knowledge - its RAG - determines an action based on what it finds, selects authorized tools, and then acts. Its path is therefore not fully written in advance: two different requests can lead to two different paths because the situation or available information is different. An agent can be triggered automatically; what is no longer fully automated in advance is its decision path - it makes a decision and then acts. This distinction - executing a process, conversing, or reasoning and deciding before acting - runs throughout this White Paper. It also explains why the AI entering the enterprise today is not coming through the same door as yesterday's automations or chatbots, and why it should not inherit their reputation.

An LLM alone is fluent in language, but it does not automatically know a company's offers, pricing, procedures, commitment limits, or history. RAG (Retrieval-Augmented Generation) provides the organization's validated content. Agentic capability interprets the objective, prepares steps, chooses authorized tools, observes the result, and adapts the path. The LLM understands and formulates; RAG provides the facts; agentic AI drives action within the defined framework.

Sacha Engine brings these three dimensions together in one central nervous system. A request can begin on the web, WhatsApp, phone, email, or, when appropriate, SMS. Context is preserved; the knowledge base is queried; business tools can be used under defined permissions; and the human team receives the conversation with its context when a handoff is needed.

This is the architecture this White Paper examines: not AI as an isolated interface, but AI connected to enterprise knowledge, rules, tools, channels, and teams.

Why this White Paper exists

The goal is not to add another collection of predictions about AI. It is to put in everyone's hands a verifiable explanation of what this technology is, what it can do, what it cannot do, and what must be built around it to produce a result.

The document is intentionally practical. It separates what is already proven from what is emerging, uses public and traceable sources, and avoids treating vendor announcements as universal facts. It explains the technology without requiring a technical background, while going far enough for an executive, CIO, sales leader, transformation leader, public-sector decision-maker, consultant, student, or entrepreneur to make informed decisions.

The central thesis is simple: AI will create enterprise value when it becomes an architecture, not merely a tool. That means connecting models to reliable knowledge, authorized actions, business systems, channels, metrics, and human judgment.

The rest of the White Paper develops this thesis in three stages: why now, how it works, and how to deploy.

WHY NOW?

The change is not only technological. It is behavioral, economic, organizational, and increasingly regulatory.

THE FIRST SHIFT IS IMMEDIACY. FOR THREE DECADES, DIGITAL SYSTEMS TRAINED PEOPLE TO SEARCH, BROWSE, COMPARE, AND WAIT FOR A RESPONSE. GENERATIVE AI IS CHANGING THAT HABIT: USERS INCREASINGLY STATE AN OBJECTIVE IN NATURAL LANGUAGE AND EXPECT THE INFORMATION, SERVICE, OR NEXT ACTION TO COME TO THEM.

The second shift is delegation. Personal AI agents are beginning to persist across tasks, use tools, remember context, and act on behalf of users. This will progressively change how customers approach companies: sometimes the customer will come directly; sometimes an agent will come first.

The third shift is authorization. As AI systems gain the ability to act, companies must define what they are allowed to access, what they may decide, what requires confirmation, and when a human must take over. Consent, traceability, governance, and security become operational design requirements.

These three shifts explain why the issue is moving from experimentation to enterprise architecture.

HOW DOES IT WORK?

MY RECOMMENDATION TO EVERYONE

Do not start with training: get equipped. Buy paid subscriptions to AI services - twenty, thirty, or fifty dollars per month per person. Free versions often provide access to less capable models with tighter usage limits; judging the technology only through them can mean judging an older state of the art.

At that price, it is not simply an expense: it is one of the most cost-effective investments you can make in learning the technology, both personally and professionally. It is also the only way to judge what the technology can do today rather than what it could do two months ago.

Be cautious about the rest. Tutorials that ask you to download "their" prompt at the bottom of a page or under a video are rarely truly free: they often exist to earn an affiliate commission or sell training. And what they teach can already be two versions behind.

Once subscribed, the best training is often the AI itself: ask it directly what it can do for you, personally and professionally. That is often more useful - and more current - than a tutorial recorded two versions earlier.

When you use it, always ask for sources and the link that lets you verify them. AI can state information without a source; requiring the source and the link lets you check the claim instead of accepting it on trust.

The technical stack is easier to understand when the roles are separated. The LLM provides language understanding and generation. RAG supplies relevant enterprise knowledge at the time of the request. Agentic AI orchestrates the objective, tools, and sequence of actions. Business systems provide the real-world data and capabilities. Human oversight sets limits and handles judgment-heavy cases.

None of these building blocks is sufficient on its own. The value comes from how they are combined, governed, measured, and connected to actual work.

HOW TO DEPLOY

Deployment begins with a business need, not with a model. The White Paper proposes a structured approach to define the use case, required knowledge, channels, tools, permissions, KPIs, human handoffs, and governance before expanding scope.

THE HOW TO DEPLOY SECTION IS DESIGNED TO TURN READING INTO A WORKING PROJECT: SEVEN STRUCTURING QUESTIONS, EIGHT DEPLOYMENT RUNGS, FOURTEEN DECISIONS TO DOCUMENT, AND A SCRIPT THAT READERS CAN PASTE INTO THEIR OWN AI TO PREPARE A FIRST VERSION OF THEIR PROJECT REQUIREMENTS DOCUMENT.

The purpose is not to automate everything. It is to determine where AI should understand, retrieve, prepare, act, or hand off - and to do so in a controlled framework.

THE GREAT SHIFT

From individual AI use to enterprise systems

The first wave of generative AI spread through individual tools. Employees opened ChatGPT, Claude, Gemini, or Copilot and discovered a new way to write, summarize, analyze, and search.

WHAT THE AVAILABLE RESEARCH MEASURES

At the enterprise level - nearly 6,000 executives surveyed between November 2025 and January 2026 in the United States, United Kingdom, Germany, and Australia: 69% of companies use AI, but nine out of ten report no measurable productivity effect over the previous three years. The average reported gain is 0.29%. Those same executives expect 1.4% over the next three years.

At the individual level - across 5,179 customer-support agents, access to a conversational AI assistant increased the number of cases resolved per hour by 14%, but by 34% for less experienced workers, with minimal effect on the most experienced.

At the team level - across 758 Boston Consulting Group consultants, AI enabled 12.2% more tasks, completed 25.1% faster, with quality more than 40% higher, and a 43% gain for those who had initially performed below average.

Where a real system has been built - 80% of technical leaders surveyed report measurable economic returns, and 57% of organizations already run agents on multi-step processes.

The next wave is different. It is not about giving every employee a smarter text box. It is about connecting AI to the organization's knowledge, systems, and workflows so that it can become part of how work is actually performed.

That transition is difficult because the problem is no longer only model quality. It becomes an architecture problem: data quality, permissions, identity, tools, process design, security, measurement, and human responsibility all matter.

This is why enterprise AI cannot be reduced to a license purchase.

A company can equip thousands of employees with ChatGPT Enterprise, Microsoft 365 Copilot, Gemini Enterprise, or Claude for Enterprise and still have no enterprise AI architecture.

Access is not integration.

WHY NOW?

The analogy with the Internet is useful. In 1996, buying an Internet connection did not create a digital strategy. Building a website did not automatically create digital acquisition, e-commerce, CRM, analytics, or customer experience. Those capabilities had to be designed around the network.

AI is entering the same phase. The model is becoming infrastructure; the competitive advantage comes from what the organization builds around it.

THE ADOPTION PARADOX

Widespread use, limited measurable impact

Multiple studies now point to the same tension: AI use is widespread, but enterprise-scale integration and measurable financial impact remain much less common.

This does not mean generative AI creates no value. It means that individual use does not automatically become an enterprise system.

The missing layer is the operating architecture around the model: enterprise knowledge, rights and permissions, business tools, channels, measurement, governance, and human oversight.

That distinction between use and integration runs throughout this White Paper.

2026 REFERENCE POINT

88% of customers expect faster responses than one year earlier; 74% expect 24/7 service availability.

Source: Zendesk, CX Trends 2026.

THE PLATFORM SHIFT

The largest technology vendors are moving from assistants toward agents

AUDIENCE	NEED AT THE MOMENT IT APPEARS	EXPECTED EXPERIENCE
Visitor	Quickly understand an offer or service	A clear, immediate answer tailored to the question
Prospect	Compare, clarify the need, resolve an objection	A useful conversation without waiting for a sales callback
Customer	Obtain information, follow-up, or a solution	Continuity of context and the right human handoff when needed
User or citizen	Access a right, procedure, or service	Available, understandable, and traceable guidance

Microsoft, Google, OpenAI, Salesforce, xAI, Meta, and others are all adding capabilities that go beyond answering questions: persistent context, tool use, workflow execution, voice, memory, permissions, and agent orchestration.

The terminology varies, and the maturity of the products varies, but the direction is consistent: models are being embedded into systems that can pursue objectives and use tools.

For enterprise leaders, this matters because the technology category itself is changing. Buying an assistant and designing an agentic system are not the same decision.

THE NEW SERVICE PROMISE

Be present at the exact moment a need appears. Immediately provide information that is precise, concise, contextualized, and verifiable. Preserve context from one channel to another. Let the user choose whether to continue, act, speak with a human, or decide when and how they want to be contacted again.

MEASURING THE GAP

Usage is not integration

The most useful way to read current adoption figures is to separate three levels: employees using AI tools; business functions piloting or deploying AI; and organizations that have connected AI to core knowledge, systems, controls, and measurable outcomes.

900 M+

Weekly ChatGPT users
OpenAI, 2026

88 %

Organizations using AI in at least one function
McKinsey, State of AI, November 2025

7 %

Have fully deployed AI across the enterprise, across all functions
McKinsey, State of AI, November 2025

94 %

Have not yet achieved enterprise-wide deployment with a significant impact on operating results
McKinsey, 2026

These levels are often mixed together in headlines, which makes the market appear more mature than it is.

Direct sources: OpenAI, "Scaling AI for everyone" · McKinsey, "AI at work but not at scale" · McKinsey, "Capturing Central Europe's AI opportunity."

These figures do not contradict one another. They distinguish use from integration. An organization can allow employees to use an assistant, experiment with a few use cases, and report that it uses AI without having connected that intelligence to its knowledge, access rights, business systems, channels, and economic indicators. The major shift of the next few years will be this move from personal tool to enterprise system.

2026-2035 · THE MARKET OPENING UP

Companies no longer want only powerful models. They want systems that produce outcomes.

The market is already investing to prepare for this shift. Gartner forecasts \$2.59 trillion in worldwide AI spending in 2026, up 47%. This figure covers all AI-related spending, including hardware and services; the same firm estimates spending specifically tied to agentic capabilities, including embedded capabilities, at \$201.9 billion in 2026 (Chapter 1). Gartner also notes that spending remains dominated by providers and hyperscalers: enterprise investment potential has not yet been fully activated. At the same time, the share of enterprise applications incorporating task-specific agents could rise from less than 5% in 2025 to 40% by the end of 2026.

95% OF PILOTS STUDIED WITH NO MEASURABLE IMPACT

The report published by MIT Project NANDA draws on more than 300 public initiatives, 52 organizational interviews, and 153 executive responses. It estimates that after \$30 billion to \$40 billion invested, 95% of the integrated pilots studied had produced no measurable P&L impact. The authors identify less a lack of model power than a lack of memory, context, learning, and process integration.

Primary source and status: preliminary findings, Project NANDA, July 2025, "The GenAI Divide: State of AI in Business 2025."

This does not mean generative AI creates no value. It shows that an individual tool does not automatically become an enterprise system. RAG provides context and verifiable knowledge; agentic AI connects that intelligence to objectives and tools; humans retain critical analysis, judgment, creativity, and decision-making. Together, these dimensions increase the organization's capacity to think, produce, analyze, and act. They are an architectural response to observed causes, not an automatic guarantee of results.

CURRENT USE	ENTERPRISE STAGE	VALUE SOUGHT
Individual assistant	Agent connected to rules and tools	An available and measurable service
Scattered documents	RAG-governed knowledge base	Reliable and verifiable answers
Separate channels	Multichannel memory and context	A seamless experience

Direct sources: Gartner, "\$2.59T in AI spending in 2026" · Gartner, "40% of applications with task-specific agents."

In Gartner's high-end scenario, agentic AI could account for more than \$450 billion in enterprise software revenue by 2035. This projection does not guarantee the success of any individual project. It indicates the market direction: value will shift from models used in isolation toward architectures able to understand a request, retrieve the right knowledge, act under control, and work with humans.

100 million prospects had never been called back at Salesforce

In August 2025, Marc Benioff, Salesforce co-founder, Chairman and CEO, revealed that more than 100 million prospects who had contacted the company during the previous 26 years had never been called back because there was not enough human capacity. Even the world's largest CRM vendor was not immune to the problem documented in this White Paper. Four months later, Salesforce announced the acquisition of Qualified, a provider of AI SDR agents. The transaction closed on April 1, 2026. In its Form 10-Q for the quarter ended April 30, 2026, Salesforce reported that the fair value of consideration transferred for Qualified was \$1.2 billion, primarily \$1.1 billion in cash (Chapter 8, Section 8.2).

Direct source: Marc Benioff, The Logan Bartlett Show, Episode 149, August 29, 2025.

READING GUIDE

Who this White Paper is for and how to use it

It is intended first for executives, sales and marketing leaders, CIOs and IT leaders, and transformation leaders evaluating how to scale generative AI. It is also relevant to HR leaders, customer-service leaders, and public-sector decision-makers, whose challenges are similar even if the vocabulary differs.

But it is not reserved for any one audience. Students, teachers, researchers, journalists, consultants, employees in any function, first-time entrepreneurs, or simply curious readers who want to understand what AI is really changing inside an organization: this document was written to be understood without technical prerequisites. Every term is defined in the glossary, every claim is dated and sourced, and every number points back to its original publication. Curiosity is the only prerequisite.

This White Paper is also for the person who does not yet have the title to own the topic but wants to be the one who brings AI into the organization: an employee who senses the company needs to move and is looking for a practical path, or a student who wants to walk into an interview with real understanding rather than a thin layer of vocabulary. And this transformation does not stop at the enterprise door. It is also entering cities, regions, and public agencies, changing the relationship between citizens and institutions in much the same way it changes the relationship between customers and companies. From the executive to the citizen, through decision-makers and employees, this White Paper was written to be useful at every level.

The document is intentionally technology-agnostic: it describes architecture patterns, public data, and documented experience before illustrating their application to a commercial use case in Chapter 8.

How to use this White Paper

The document can be read from beginning to end, but it can also be used as a decision tool depending on the reader's role.

NEED	RECOMMENDED PATH
Discover the topic with no prerequisites	Founder's opening, 2017-2026 milestones, then the glossary in Appendix B
Understand the concepts	2017-2026 milestones, then Chapters 4 through 7
Assess the market	Chapters 1, 6, and 7, with the Israel and international focus
Develop sales action and service	Chapter 8: pre-sales, after-sales, and multichannel
Prepare a deployment	Chapters 9 through 11, then the 14-decision grid, pages 87 and 88
Arbitrate and decide	Conclusion, then final decision summary

Table 2: Six entry points depending on what you are looking for. None requires prior technical expertise.

THREE SHIFTS ARE CHANGING THE RULES

My view of the new relationship between people and organizations

Since 1996, I have worked on the same question: how does a company meet demand at the right moment? The Internet first gave people a new power: the ability to search for themselves. Thirty years later, AI opens another stage. We no longer want only to search. We state a need and expect the information, service, or action to come to us.

This shift may look subtle because it is settling into everyday life. Yet it is profoundly changing our tolerance for delay. Ordering a meal, booking a ride, making a payment, translating text, comparing options, or summarizing a file can now be done in a few taps - sometimes in a single sentence.

And that is where I see a paradox. We can obtain a complex synthesis in seconds, yet we still often have to fill out a form to find out whether a car is available, whether an insurance policy fits our situation, whether a property meets our criteria, or simply to get information from an organization. Technology has changed faster than service delivery.

MY CONVICTION

The real shift is not only: "AI answers better." It is: "I state my need, and I expect the system to immediately organize the right answer, the right service, or the right action."

These three shifts define the new conditions for action: act at the right moment, through the right intermediary, and with the right to do so.

01

Immediate action - the real-time requirement

Expectations form and decisions happen in the moment. An enterprise that responds later loses the value of intent.

02

Delegated action - personal agents are arriving

Soon, each person may have an agent that understands, anticipates, and acts on their behalf. The interface shifts from the website or app toward the personal agent.

03

Authorized action - law and consent change the rules

Transparency, consent, and control become the norm. The ability to act with customer authorization becomes a competitive advantage.

These three movements tell the same story: value is moving toward the moment intent appears. Companies that can be useful, legitimate, and immediate at that moment will win the relationship.

Customers are no longer lost because of visibility. They are lost because of inaction.

This is the lens I propose before the technical chapters. Agentic AI, RAG, LLMs, channels, and governance are not a collection of technologies; they are the building blocks of a new enterprise availability.

WHAT IS THIS ANALYSIS BASED ON?

Immediacy becomes a norm - Zendesk, CX Trends 2026.

Personal agents are arriving - Google I/O 2026, "Welcome to the agentic Gemini era"; Meta, "The Future is for Everyone"; Microsoft, "Human-led, agent-operated."

Consent and evidence become structural - Légifrance, Article L223-1 of the French Consumer Code; FCC, "AI-generated voices in robocalls."

Full sources and links: appendix, pages 127 to 129.

The end of delayed sales cycles. The rise of real-time selling.

The Internet first revolutionized access. A person could find information that previously required travel, a phone call, a catalog, an intermediary, or time. We learned to type a query, open links, browse pages, and compare.

The smartphone and on-demand economy then revolutionized speed. The gesture became shorter: I tap a screen and a ride arrives, a meal is prepared, a payment is sent, a package is tracked. Value also comes from removing unnecessary waiting and making the path from request to result visible.

Generative AI adds a third movement. Instead of searching several pages and building an answer myself, I state the objective in natural language. The system searches, connects, summarizes, and explains. Agents add another step: they can continue after the answer, use tools, and prepare an action. Google describes Gemini Spark as a personal agent available 24/7; OpenAI is evolving ChatGPT toward long-running, multi-step tasks; Microsoft speaks of "human-led, agent-operated" organizations.

In 2026, Zendesk measured a signal consistent with this movement: 88% of respondents expect faster responses than a year earlier, and 74% associate AI with always-on availability. This is not a universal law of behavior, but it is a clear indicator of what AI is beginning to make normal.

YESTERDAY	TODAY	TOMORROW
I search	I ask	I delegate part of the search or action
I browse pages	I receive a synthesis	An agent monitors, compares, alerts, or prepares the next step
The website is the interface	The conversation becomes the interface	The agent can become the intermediary

Table 3: Three eras of access to information and service.

The Internet gave us access. Mobile gave us immediacy. AI gives us conversation. Agentic AI is beginning to give us delegation.

Sources for this page: Zendesk, CX Trends 2026 · Google I/O 2026 and Google, "The Gemini app becomes more agentic" · Microsoft, "Human-led, agent-operated" · OpenAI, ChatGPT Work and Workspace agents. Full references and links: appendix, pages 127 to 129.

SHIFT 01 · THE FORM IS CHANGING ITS ROLE

A form records demand. It does not understand it.

For years, the form has been the default bridge between digital traffic and a sales team. It asks for a name, email address, phone number, and sometimes a few qualifying fields. It is efficient for storage, but weak for understanding.

A form cannot ask a follow-up question because it detected hesitation. It cannot explain a difference between two offers. It cannot adapt to urgency. It cannot detect that the user has misunderstood a condition. It cannot decide that the request is ready for a human right now.

This matters because intent has a half-life. The person who just clicked an ad, scanned a QR code, opened WhatsApp, or called a number is expressing a need in the present. The longer the organization waits, the more context is lost and the more likely the user is to move elsewhere.

The question is therefore not whether forms disappear. They will remain useful. The question is what should happen before, during, and after the form when AI can understand the request in real time.

The form can become one data-capture step inside a larger conversation instead of the conversation's endpoint.

That changes the design principle: do not begin with "What fields do we need?" Begin with "What does the person need, and what must the system understand before it can help?"

MY CONVICTION

The form does not disappear. It simply should no longer force a person to wait before explaining what they want.

Name, email, phone: fine. But the most important question becomes: "What do you need?"

And tomorrow, that request may not even be formulated directly by the person: their own AI agent may begin preparing it for them.

SHIFT 02 · PERSONAL AGENTS ARRIVE ON THE CUSTOMER SIDE

Your customer may send an AI agent to talk to your company

Consider an already familiar request: "I need renters or homeowners insurance for this property, with this coverage and this budget." Today, the person opens several websites, fills out forms, waits for quotes, and compares them manually. Tomorrow, part of that work can be delegated to a personal agent.

The agent can know criteria the person has agreed to share, monitor the market, eliminate incompatible offers, and ask for missing information. The same scenario applies to a vehicle, a trip, a contractor, a property, or a local service.

The question then shifts to the other side: what does the agent find when it reaches the company? If it finds only a marketing page and a form, it lacks operational facts: current pricing, availability, rules, required documents, timing, appointment slots, commitment limits, and the conditions under which a human must step in.

EDITORIAL PROJECTION

Tomorrow, an enterprise will not receive only people. It will also receive requests prepared or transmitted by the agents working for them.

I am not claiming that all agents will automatically negotiate with one another tomorrow. That would go beyond the available evidence. But the necessary building blocks are already emerging: persistent personal agents, tools, voice, mandates, permissions, memory, and enterprise agents.

Table 4: What changes when an agent appears on the customer side.

A POWERFUL PERSONAL AGENT IS OF LITTLE USE IF THE COMPANY ON THE OTHER SIDE CANNOT EXPOSE RELIABLE KNOWLEDGE, RESPOND, AND ACT PROPERLY.

SHIFT 03 · LAW AND CONSENT ARE CHANGING THE RULES

When a customer asks to be contacted, the company must be able to prove it

The third shift is regulatory. Rules differ by country, channel, and industry, so shortcuts should be avoided. But the direction is clear: a company must know why it is contacting someone, on what legal basis, for what purpose, and with what evidence.

France, August 11, 2026. The French Consumer Code shifts B2C telephone solicitation toward prior consent. The business must be able to demonstrate that consent.

For certain sectors, including energy-efficiency renovation and home adaptation, the July 23, 2026 decree provides a very concrete example of the value of an explicit inbound request: when a consumer asks to be called back, the company must be able to substantiate that request, respect its scope, and act within the required period - including five business days in the case covered by the decree.

WHAT IS CONVERGING

The interfaces differ: search engine, messaging, office suite, phone, operating system, commerce, voice assistant. The direction is the same: understand context, use tools, pursue a goal.

Channels must also be distinguished. August 11, 2026 does not create a new blanket ban on all emails and SMS messages: for B2C electronic marketing in France and the EU, prior consent was already the general principle, subject to defined exceptions.

IN THE UNITED STATES, THE FCC HAS CONFIRMED THAT AI-GENERATED VOICES FALL WITHIN TCPA RULES GOVERNING ARTIFICIAL OR PRERECORDED VOICES. HERE TOO, THE DETAILS DEPEND ON THE TYPE AND PURPOSE OF THE CALL.

In this context, an inbound request has a source, date, channel, purpose, sometimes consent, and sometimes a time limit. A serious agentic architecture must be able to preserve those elements and enforce them throughout the interaction.

Legal sources: Légifrance, Article L223-1 of the French Consumer Code; Légifrance, Decree No. 2026-662 of July 23, 2026; CNIL, commercial prospecting by email, SMS, MMS and automated calling; FCC, "AI-generated voices in robocalls." Full references: appendix, pages 127 to 129.

CHANNELS · WHERE DEMAND APPEARS

Web, WhatsApp, voice: three entry points to the same need

The channel is no longer just a medium. It is where intent appears. A request can come from a Google search, a click on a Meta ad, a WhatsApp message, or a phone call. Three entry points are especially strong for immediate conversation: web, WhatsApp, and voice. Email and SMS play more of a continuity, documentation, and confirmation role.

The web remains central, but AI can reverse an old logic: instead of forcing users to translate their question into navigation, they can ask it in their own words and the system can retrieve the appropriate information from enterprise knowledge.

THE THROUGHLINE

The serious debate is no longer only about the quality of an answer. It is about the distribution of action: what can the agent do, based on what knowledge, within what limits, and at what point does a human decide?

WhatsApp is different because conversation is already the interface. Meta reports more than 3 billion users across more than 180 countries, and more than one billion active threads with businesses every day across WhatsApp, Messenger, and Instagram. An ad can therefore lead directly to a conversation without forcing an intermediate page.

VOICE RETAINS PARTICULAR POWER: WHEN SOMEONE CALLS, THEY ARE USUALLY EXPRESSING IMMEDIATE INTENT. BUT HUMAN-OPERATED VOICE IS EXPENSIVE BECAUSE IT REQUIRES SYNCHRONOUS TIME, STAFFING HOURS, LANGUAGES, AND AVAILABILITY. XAI, OPENAI, AND OTHERS ARE NOW INVESTING IN AGENTS CAPABLE OF CONDUCTING VOICE CONVERSATIONS, USING TOOLS, AND HANDING OFF TO A HUMAN WHEN NEEDED.

Table 5: Each channel has its own role, but one knowledge layer should sit behind all of them.

Multichannel does not mean being everywhere. It means understanding the same request regardless of where it enters the enterprise.

Sources for this page: Meta, Business Agent and "WhatsApp message translations"; xAI, Voice API and Grok Voice Agent Builder; OpenAI, Operator. Full references and links: appendix, pages 127 to 129.

WHAT THESE THREE SHIFTS CHANGE FOR THE ENTERPRISE

WHAT I TAKE FROM THIS

The Otto signal is interesting not because it announces the end of the cloud, but because it makes one idea tangible: the agent is no longer just a conversation open in a browser tab. It becomes a process that remains present, observes, executes, and waits for the next instruction.

Visibility is no longer enough. You must respond when demand arrives.

For thirty years, digital marketing mainly perfected the top of the funnel: visibility, discoverability, paid traffic, click measurement, and conversion. These disciplines remain essential. But they are no longer sufficient when users expect an immediate response and when an agent can begin working on their behalf.

I THEREFORE PROPOSE ADDING ONE QUESTION TO THE CLASSIC ACQUISITION QUESTIONS: WHAT HAPPENS IN THE SECONDS AFTER THE NEED APPEARS? IF THE ANSWER IS "A FORM AND A CALLBACK LATER," THE COMPANY IS STILL OPERATING WITH AN ORGANIZATION INHERITED FROM YESTERDAY'S WEB. IF THE ANSWER IS "A RELIABLE, CONTEXTUALIZED, TRACEABLE CONVERSATION CAPABLE OF PREPARING ACTION," THE COMPANY IS BEGINNING TO ADOPT THE NEW LENS.

This new lens does not require total autonomy. On the contrary, major technology vendors are all adding permissions, controls, confirmation, and human oversight. Google places Spark under user direction. Microsoft speaks of human-led organizations. OpenAI builds agents in controlled environments. xAI emphasizes guardrails and tools. Meta lets the business choose when a human takes over.

That is exactly the model I advocate: humans keep the final say, but they no longer need to be the first link in every interaction. AI can handle immediacy, knowledge retrieval, qualification, and preparation for action. Humans step in where judgment, empathy, expertise, negotiation, or accountability make the difference.

Yesterday, the Internet taught us to search. Today, AI is teaching us to ask. Tomorrow, agents will act for us. The enterprise must be ready when demand appears.

Sources for this page: Google I/O 2026; Meta Business Agent; Microsoft, "Human-led, agent-operated"; OpenAI, ChatGPT Work; xAI, agentic tools. Full references and links: appendix, pages 127 to 129.

BEFORE THE CHAPTERS BEGIN

TODAY'S JOURNEY	PLAUSIBLE AGENTIC JOURNEY
The person searches across several sites	The person states their criteria once
They reconstruct the information themselves	Their agent brings together the accessible information
She fills out several forms	The agent asks only for what is missing, within the authorized framework
They wait for a callback	The enterprise system can respond immediately or organize the right human handoff

Four ideas to keep in mind

This opening does not replace the chapters that follow. It provides a lens for reading them. To avoid repeating the same arguments, the following pages develop each theme where it becomes operational.

TABLE 6: FOUR RECURRING THEMES AND WHERE EACH BECOMES OPERATIONAL.

The technical chapters therefore keep their role: they explain LLMs, RAG, agentic AI, channels, governance, security, and the human-complementarity model. The opening explains why these building blocks are becoming strategic now.

PART I

The enterprise AI paradox, the platforms shifting toward agents, and the numerical evidence.

Three years after language models became widely available, the picture is strikingly asymmetric: nearly every organization reports using AI, yet relatively few see a measurable effect on their P&L. This chapter examines that gap, its documented causes, and what it means for senior leadership.

1.1 Adoption has become widespread

McKinsey's global survey published in late 2025 found that 88% of surveyed organizations regularly use AI in at least one business function, up from 78% the previous year, and that 72% use generative AI, up from 33% in 2024. Interest in agents is high: 62% of respondents report at least experimenting with AI agents.

WHAT THIS CHANGES

The strategic point is not: "all prospecting will disappear." The point is that proof of intent and of the right to continue the relationship becomes more valuable.

These headline numbers are important, but they do not tell us how deeply AI is integrated into the organization.

The next question is where the technology is actually embedded: isolated personal tools, departmental pilots, or systems connected to enterprise knowledge and workflows.

1.2 THE PRODUCTIVITY PARADOX

High adoption does not automatically produce measurable enterprise productivity.

Several studies have highlighted the distance between experimentation and scaled impact. The reasons include fragmented use cases, weak process redesign, poor data foundations, lack of governance, limited integration with business systems, and unclear metrics.

The issue is not whether AI can improve individual tasks. It clearly can. The issue is whether those improvements accumulate into enterprise-level outcomes.

That requires architecture and operating change around the model.

1.3 What executives should measure

CHANNEL	PRIMARY ROLE	WHAT MUST BE RETAINED
Web	Understand intent during the visit	Page, source, campaign, question, useful context
WhatsApp and messaging	Start a conversation directly	Thread, origin, information provided, consent framework
Phone and voice	Handle a synchronous request	Call origin, purpose, right to call back, trace
Email and SMS	Document, notify, confirm	Purpose, history, sending basis, and opt-out where used for marketing

AI programs should not be evaluated only through usage metrics such as number of prompts, active users, or licenses distributed.

The relevant indicators depend on the business case: response time, conversion, cost per interaction, resolution rate, time saved, error rate, revenue influenced, customer satisfaction, or employee throughput.

A strong metric connects AI activity to an operational or financial outcome.

THE FIRST LEADERSHIP DECISION IS THEREFORE NOT "WHICH MODEL SHOULD WE BUY?" IT IS "WHICH OUTCOME ARE WE TRYING TO CHANGE, AND HOW WILL WE KNOW?"

Table 7: From adoption metrics to business-outcome metrics.

1.4 The architecture gap

The largest gap between experimentation and scaled value often appears between the model and the rest of the enterprise.

The model may be capable, but it cannot create reliable outcomes if it lacks access to the right knowledge, cannot use the right tools, does not know its permissions, or is not connected to human workflows.

This is the core architecture problem the next chapters address.

01	Capture Source, channel, date, purpose, proof of request or consent where required.
02	Understand Intent, context, urgency, objection, missing information.
03	Know Enterprise knowledge validated and governed through RAG.
04	Act Authorized tools, controlled steps, explicit limits.
05	Continue Context preserved across web, WhatsApp, voice, email, and SMS.
06	Hand off Human informed, full context, decision at the right moment.

PART I · WHY NOW?

01. The enterprise AI paradox

02. THE PLATFORMS SHAPING ENTERPRISE AI

03. Customer cases and quantified evidence

The point is not to predict that every company will become "agentic." The point is to understand why the technology, the market, and user expectations are converging around systems that can understand, retrieve, and act.

THROUGH-LINE	WHERE IT REAPPEARS IN THE WHITE PAPER	WHAT TO READ THERE
Immediacy	Pages 7-8; Chapter 8.1; Chapter 8.5	The need appears now; speed becomes a component of experience and commercial value.
Personal agents	Chapter 6; architecture in Chapter 9	The enterprise must remain interoperable, expose reliable knowledge, and preserve action rules.
Consent and inbound demand	Chapter 8.6; Chapter 10	Provenance, evidence, channel, and the right to continue the relationship become operational data.
Human at the center	Chapters 3, 5, 9, and 10	Autonomy is bounded; humans arbitrate, negotiate, approve, and retain responsibility.

The rest of Part I documents that convergence.

READING RULE

The editorial principle is simple: one strong idea in the opening; one proof point, architecture, or use case where the chapter needs it. No full repetition.

Source names are shown directly on the relevant pages. Complete references, dates, and verification links are grouped in the appendix, pages 127 to 129. The appendix is there for verification; the page remains understandable without leaving the reading flow.

CHAPTER 1

CROSS-CUTTING RULE - THE ENTERPRISE MUST EVOLVE AT THE SAME PACE AS AI

The chapters that follow should be read with a fifth idea in mind: tools, uses, and development methods are evolving in parallel. The right strategy is therefore to maintain an overall architectural vision while building in stages around real needs, incorporating new advances at each rung.

THE ENTERPRISE AI PARADOX

WHY NOW?

Use is widespread. Measurable impact is not.

01

The enterprise AI paradox

*Usage is expanding everywhere.
Financial performance remains concentrated.*

The first chapter separates enthusiasm from evidence and shows why the enterprise question is no longer whether to use AI, but how to turn it into a governed operating capability.

WHAT CONVERGES AS OF AUGUST 19, 2026

The LLM becomes a selectable layer; RAG becomes multimodal, evaluated, and governed; agentic AI becomes an orchestration layer with identity, permissions, memory, tools, and traces; standards finally make it possible to connect these building blocks without locking the enterprise into one provider.

The adoption curve moved faster than the integration curve.

That difference is the starting point for the rest of the White Paper.

Source: McKinsey & Company, "The state of AI in 2025: Agents, innovation, and transformation," November 2025.

1.2 Transformation is still uncommon

The same research shows that nearly two-thirds of organizations remain in experimentation or pilot mode, while roughly one-third have begun to scale. Financially, 39% of respondents attribute some operating-income impact to AI, and an estimated 6% make up the group of "high performers" that attribute more than 5% of EBIT to AI.

The preliminary field report "The GenAI Divide: State of AI in Business 2025," published by MIT's Project NANDA, reviews more than 300 public initiatives and draws on 52 organizational interviews and 153 executive responses. After \$30 billion to \$40 billion in investment, the authors estimate that 95% of the integrated pilots they observed had produced no measurable P&L impact. Their diagnosis focuses less on model power than on missing context, memory, learning, and process integration. This signal is not, by

itself, a scientific consensus; it powerfully documents the gap between an impressive demo and an enterprise system capable of producing value.

Primary source, preliminary findings: Project NANDA, "The GenAI Divide: State of AI in Business 2025," July 2025.

THE PROBLEM STATEMENT

Organizations that achieve results share three characteristics: they ground the model in their own data, redesign the process around the agent instead of adding the agent to an existing process, and measure quality continuously.

The rest of the document explains precisely how these three conditions translate into architecture.

1.3 The market is already anticipating this shift

Forecasts differ on size but converge on direction. Gartner places worldwide agentic-AI spending at \$201.9 billion in 2026 when agentic capabilities embedded in enterprise software are included, and forecasts that 40% of enterprise applications will include a task-specific agent by the end of 2026, up from less than 5% one year earlier. Firms that measure the stand-alone agentic-AI market - Fortune Business Insights, Precedence Research, MarketsandMarkets, and Deloitte - estimate it at \$7 billion to \$12 billion for 2025-2026, with compound annual growth rates between 38% and 46%.

The same firm issues the corresponding warning: more than 40% of agentic-AI projects could be canceled by the end of 2027, mainly because of poorly defined value, uncontrolled costs, and insufficient governance. The two forecasts should be read together: the market is growing quickly, and selection will be unforgiving.

FIRM	MEASURED SCOPE	ESTIMATE	HORIZON
Gartner	Agentic spending, including embedded capabilities	201,9 Md\$	2026
Fortune Business Insights	Standalone agentic AI market	7,29 Md\$ → 139,19 Md\$ (TCAC 40,5 %)	2025 → 2034
Precedence Research	Standalone agentic AI market	7,55 Md\$ → 199,05 Md\$ (TCAC 43,8 %)	2025 → 2034
MarketsandMarkets	Standalone agentic AI market	7,06 Md\$ → 93,20 Md\$ (TCAC 44,6 %)	2025 → 2032
Deloitte	Standalone agentic AI market	8,5 Md\$ → 35-45 Md\$	2026 → 2030
Next Move Strategy Consulting	RAG market	3,33 Md\$ → 81,51 Md\$ (TCAC 42,7 %)	2026 → 2035
MarketsandMarkets	RAG market	1,94 Md\$ → 9,86 Md\$ (TCAC 38,4 %)	2025 → 2030
The Business Research Company	AI SDR software	5,81 Md\$ → 17,58 Md\$ (TCAC 32,3 %)	2026 → 2030

Table 7: Estimates published between 2025 and 2026. The differences come from market definitions, not disagreement about the direction of travel.

Sources: Gartner, 2026-2027 forecasts; Fortune Business Insights; Precedence Research; MarketsandMarkets; Deloitte, TMT Predictions; Next Move Strategy Consulting; The Business Research Company.

A detailed timeline of this development - ten years of research, products, and standards, from the foundational 2017 paper to managed agents in 2026 - appears in Appendix F so that this section can remain focused on the question that grounds it: why now.

02

The platforms shaping enterprise AI

When six competitors ship the same architecture within eighteen months, it is no longer a vendor hypothesis: it is a de facto standard.

Between 2025 and 2026, the major companies shaping enterprise AI all delivered the same combination: an agentic reasoning engine, persistent memory, an indexed knowledge layer, a standardized tool catalog, and governance and evaluation capabilities. Different commercial strategies, but the same architectural pattern. This chapter compares those building blocks and draws the practical implications of that convergence for buyers.

2.1 Amazon and AWS: industrializing the agent

Amazon has made production deployment the centerpiece of its argument. Amazon Bedrock AgentCore is positioned as the platform designed to move agents from demonstration to operation. Metrics reported in 2026 show traction: the number of tasks executed by agents on AgentCore increased fifteenfold in six months, and the SDK exceeded two million downloads five months after it was highlighted.

Announcements at the AWS Summit in New York in June 2026 describe precisely the architecture presented in Chapter 7:

- Fully Managed Knowledge Bases: a managed enterprise RAG pipeline with native connectors, managed preparation of heterogeneous formats, and an "Agentic Retriever" dedicated to complex, multi-step queries.
- AWS Context: a knowledge graph designed to tell agents where to find the information needed to answer accurately or decide the next action.
- Web Search on AgentCore: a managed tool that grounds answers in current, cited web information without moving data outside the customer's AWS environment.
- Harness, Policy, Evaluations, and Payments: respectively, accelerated agent assembly in two API calls, deterministic controls over what the agent is allowed to do, continuous quality evaluation on production traffic, and autonomous transaction execution.

Andy Jassy summarized the logic in his Q2 2026 earnings communication: reliably assembling these building blocks is the real difficulty, and it is precisely what has blocked many production deployments; AgentCore

exists to provide them as managed infrastructure. He added an important forecast: while companies will build custom agents of their own, most will also use ready-made agentic services.

AWS customer examples include the PGA Tour, whose multi-agent content-generation system produces tournament coverage ten times faster, as well as Nasdaq, Visa, and Experian for enterprise-scale agent deployments.

Sources: Amazon, "AWS Summit New York 2026," June 2026; AWS News Blog; Amazon, Andy Jassy Q2 2026 earnings communication.

2.2 Jeff Bezos: the bet on industrial AI

Jeff Bezos's return to an operating role is itself a market signal. In November 2025, he co-founded Prometheus with Vik Bajaj, a former Google X executive and Verily co-founder, and became co-CEO - his first executive role since stepping down as Amazon CEO in 2021. The company launched with \$6.2 billion in funding.

In June 2026, Prometheus announced a \$12 billion Series B at a valuation of roughly \$41 billion, with investors including JPMorgan Chase, Goldman Sachs, and BlackRock. Its stated goal: build an "artificial general engineer," a system capable of helping design and manufacture complex physical products, from aircraft engines to pharmaceutical compounds. Bezos has publicly described the ambition as a very modern version of computer-aided design and has specified that the project is not about robotics. The company operates from San Francisco, London, and Zurich.

Sources: CNBC, June 11, 2026; TechCrunch, June 11, 2026; Axios, June 11, 2026; GeekWire, June 11, 2026.

THE STRATEGIC READING

Bezos's bet focuses on AI applied to the physical world. Its significance for this White Paper lies in the shift it illustrates: value is moving from conversation to execution. A system that designs an engine and an agent that runs a complete sales cycle are part of the same underlying movement: AI that produces an outcome rather than AI that produces an answer.

2.3 Microsoft, Google, OpenAI, Anthropic, xAI, Salesforce, NVIDIA

ACTEUR	AGENTIC OFFERING	POSITION AND DIFFERENTIATION
Microsoft	Copilot Studio, Azure AI Foundry, Agent 365	Identity inheritance: agents reuse existing Entra ID, Purview, and Defender policies. Predictable per-user billing. A decisive advantage in organizations standardized on Microsoft 365.
Google	Gemini Enterprise, Vertex AI Agent Builder and Agent Engine, Antigravity	Very broad model choice through Model Garden, sessions and Memory Bank for persistent infrastructure-level memory, unified governance and audit trail.
OpenAI	AgentKit, Agents SDK, Workspace Agents	Developer-tool approach: build, evaluate, and deploy quickly. Broad connector ecosystem, persistent memory by workspace.
Anthropic	Claude, Model Context Protocol	The interoperability protocol that became a de facto standard, transferred in December 2025 to the Agentic AI Foundation under the Linux Foundation.
xAI	Grok API, agentic tools, function calling	A reasoning-model family accessible by API, with tool calls, structured outputs, and integration into a multi-model architecture.
Salesforce	Agentforce 360, Atlas Reasoning Engine, Data Cloud	Agents grounded in CRM data. Published metric in units of agentic work rather than licenses: a shift in language from seats sold to work performed.
AWS	Bedrock AgentCore, Knowledge Bases, AWS Context	Managed agentic infrastructure, framework-agnostic, with built-in evaluation and policies.
Specialized vendors	Sierra, Decagon, Qualified, Intercom Fin	Vertical customer-contact agents, often priced per resolution: the business model aligns with outcomes.
SACHA OMON AI	Sacha Engine (agentic engine) and dedicated RAG for each customer	A single multichannel agentic AI SDR agent across web, WhatsApp, phone, email, and SMS, with centralized context; single-tenant infrastructure; accounts and data in the customer's name.

Table 8: Overview of agentic offerings shaping the U.S. market, with SACHA OMON AI included as a positioning reference; status as of mid-2026.

2.4 Three things the major platforms now have in common

1. Enterprise knowledge is becoming an infrastructure product. Managed Knowledge Bases at AWS, Data Cloud at Salesforce, Memory Bank and Model Garden at Google: document grounding has moved from an application component to a platform service.
2. Persistent memory is becoming a production requirement. The three largest model providers delivered long-term memory mechanisms in 2026. This directly overlaps with the failure cause identified by MIT.
3. Governance determines whether deployment is viable. Deterministic policies, continuous evaluation on real traffic, logging, and agent identity now appear in every platform catalog. The market has internalized that autonomy without traceability is not sellable in the enterprise.

WHAT THIS CONVERGENCE MEANS FOR A BUYER

When six direct competitors deliver the same architecture within eighteen months, the pattern stops being a vendor hypothesis and becomes a de facto standard. Three practical consequences follow.

The decision shifts from the model to operations. Reasoning engines can be substituted; knowledge-base quality, orchestration precision, and operating rigor create the durable difference.

Portability becomes a contractual criterion. Adoption of an open, collectively governed protocol can be verified, requested, and written into a contract.

Proof happens on real data. Major platforms have delivered continuous-evaluation systems precisely because a demo no longer predicts production performance.

2.5 The next stage: AI agents begin talking to one another

The protocols described in this chapter do more than connect an enterprise agent to its own tools. A second layer of standards, deployed by the same major players between 2025 and 2026, aims to connect agents to one another and points toward a shift in commerce that marketing and sales leaders should anticipate.

Google released the Agent2Agent (A2A) protocol on April 9, 2025, then transferred governance to the Linux Foundation on June 23, 2025. Where MCP connects an agent to tools, A2A connects agents to one another: mutual discovery, secure information exchange, and coordinated actions. When version 1.0 was released in April 2026, the protocol had more than 150 partner organizations, including AWS, Cisco, IBM, Microsoft, Salesforce, SAP, and ServiceNow.

On September 16, 2025, Google extended this foundation with the Agent Payments Protocol (AP2), an A2A extension designed with more than sixty payment partners - Mastercard, American Express, PayPal, Adyen, Coinbase - and commerce companies including Salesforce and Etsy. AP2 represents each purchase executed by an agent through three cryptographically signed "mandates": the user's stated intent, the cart assembled by the agent, and the authorized payment. The goal is to provide verifiable proof that a real person authorized the action the agent executes on that person's behalf.

OpenAI, meanwhile, launched the Agentic Commerce Protocol (ACP) with Stripe on September 29, 2025. ACP powers ChatGPT's Instant Checkout, expanded on February 16, 2026 to all U.S. users under the name "Buy it in ChatGPT." With more than 900 million weekly active users, ChatGPT becomes a sales channel in its own right, alongside a website.

FOUNDER'S FORWARD-LOOKING VIEW - HYPOTHESIS, NOT PREDICTION

This forward-looking interpretation is proposed by Salomon Bellaïche. It is based on real protocols and research, but it describes neither a stabilized standard nor a certain trajectory. The agentic web may evolve differently.

DIAGRAM 1

The agentic web

The personal agent discovers the enterprise agent, talks to it, and can then complete a transaction.

OPEN AGENT REGISTRY - PUBLIC REGISTRIES, AGENT DIRECTORIES, GPTS • DISCOVERY	
A	PERSONAL AGENT - USER SIDE Agent that accompanies an individual: preferences, budget, history.
B	ENTERPRISE AGENT - ENTERPRISE SIDE Built on business rules, products, pricing, and availability.
1	DIALOGUE — PROTOCOLE A2A Questions, needs, criteria, availability, information.
2	TRANSACTION — PROTOCOLE AP2 Mandate, agreement, negotiation, reservation, purchase.

Answer obtained or purchase completed without a human ever visiting the company's website.

Figure 1: In this vision, a personal agent discovers a company's AI agent through an open registry, communicates with it, and can proceed all the way to a transaction.

This hypothesis converges with several building blocks already visible. At Google I/O 2026, Google introduced a unified shopping cart that turns Gemini into a personal shopping agent across its services, as well as a Business Agent that lets a customer talk directly with a brand's AI agent instead of only visiting its website.

In parallel, MIT Media Lab's Project NANDA describes a web that is more readable by AI agents: an open directory, the NANDA Index, and a verifiable identity format, AgentFacts, so that one agent can authenticate another before exchanging information.

My reading of this convergence is as follows: tomorrow, a customer's personal agent could query a company's AI agent directly to obtain information, compare an offer, or complete a purchase. That is a plausible direction, not a certainty. For companies, the pragmatic point is already valid today: building an accurate AI agent, documented through RAG and capable of communicating through interoperable protocols prepares for several possible futures without presuming any one of them.

Sources: Google I/O 2026 · Google Cloud · MIT Media Lab, Project NANDA · OpenAI, AP2 and ACP, 2025-2026.

2.6 Israel: when major corporations buy, fund, and build AI

The signal to watch is not only the number of startups. It is where major corporations acquire technology, hire teams, and establish lasting research and development capacity.

85 Md\$

Israeli high-tech exports in 2025
Israel Innovation Authority, 2026

≈ 15 Md\$

Capital raised in 2025
Israel Innovation Authority, 2026

84 Md\$*

Exits high-tech 2025-2026
Israel Innovation Authority, 2026

35 %

Investment directed toward core AI
Israel Innovation Authority, 2026

** The exit figure covers Israeli high tech overall, not AI alone: the largest 2025 transactions notably included cybersecurity. It measures the depth of the ecosystem and the appetite of global buyers.*

2025-2026: AI signals to watch

TRANSACTION	KEY TAKEAWAY
Apple and Q.AI	Acquisition confirmed in January 2026. Reuters estimates the value at around \$1.6 billion. Q.AI works on audio AI.
Anthropic and Decart AI	Talks disclosed on August 13, 2026 around a potential transaction worth up to approximately \$6 billion. As of the date of this White Paper, the acquisition has not been finalized.
NVIDIA	Run:ai, an Israeli startup specializing in AI infrastructure optimization, was acquired for approximately \$700 million in late 2024. In 2025, NVIDIA launched an expansion project in Israel that could accommodate several thousand employees; later plans mention up to 10,000.
Meta and SSI	Meta attempted to acquire Safe Superintelligence, which has a presence in Tel Aviv, then hired its co-founder Daniel Gross after the offer was rejected. This is a signal of the talent war, not an acquisition of SSI.
Amazon - structural reference point	Annapurna Labs is not a 2025-2026 acquisition: Amazon bought it in 2015. But this Israeli team has become a core building block of AWS chips, including Trainium. It is a historical precedent, not a recent deal.

Table 9: Israeli transactions that document major players' appetite for AI.

Verification: AI21 Labs is not counted as an NVIDIA acquisition. Discussions were reported in late 2025, and AI21 officially disputed the existence of acquisition talks in January 2026.

France and Italy: a different profile

COUNTRY	OBSERVED PROFILE
Israel	Deep-tech density, foreign capital, and exits to global acquirers. Acquisitions and the talent war are strong signals.
France	EUR 109 billion in announced AI investment commitments in 2025, primarily for infrastructure and data centers. ASML invested EUR 1.3 billion in Mistral AI; Mistral acquired Koyeb in 2026.
Italy	Approximately EUR 10 billion in data-center investments announced for 2025-2026. G42 and iGenius launched a major AI supercomputing project; the visible signal is more about infrastructure and industrial use than AI startup M&A.

Table 10: Three ecosystems, three investment profiles. Illustrative, non-exhaustive comparison.

Reading: in the sources reviewed, Israel stands out primarily for the density of acquirable technologies and the appetite of international buyers. France concentrates more large infrastructure and financing projects; Italy more computing capacity and industrial applications.

Sources verified as of August 19, 2026: Israel Innovation Authority, High-Tech Status Report 2026; Reuters (Apple/Q.AI, January 29, 2026; Anthropic/Decart, August 13, 2026; NVIDIA, 2024-2025; Meta/SSI, July 3, 2025; France, February 9 and September 9, 2025, and February 17, 2026; Italy, January 16 and May 16, 2025).

2.7 United States, Israel, Europe, Asia: one foundation, four adoption logics

DIAGRAM 2

One global market, four adoption logics

A common architecture, adapted to the logic of each region.

UNITED STATES	ISRAEL	EUROPE	ASIA
Scale and revenue	Startup Nation	Trust and control	Channels and languages
Rapid deployment, CRM integration, pipeline measurement, and commercial ROI.	Research, talent, entrepreneurship, applied innovation, and international ambition from day one.	Governance, compliance, data sovereignty, and explicit human oversight.	Mobile-first experiences, messaging, voice, always-on availability, and linguistic diversity.

One agentic architecture, four buying logics.

Figure 2: The same agentic architecture is deployed according to different regional logics.

In the United States, buyers expect proof of revenue contribution, rapid integration into the commercial stack, and the ability to scale. In Israel, the ecosystem values technical depth, applied innovation, robustness, and international ambition. In Europe, decisions are heavily shaped by governance, localization, data ownership, and human oversight. In Asia, language diversity, mobile usage, messaging, and availability across multiple time zones become central.

These expectations complement one another and form a global design framework: produce measurable value, innovate quickly, build trust, and adapt to real-world usage. France retains a role in this picture, particularly for compliance, language, and certain industries, but it should not define the size or ambition of the addressable market on its own.

REGION	DOMINANT BUYING QUESTION	ARCHITECTURAL RESPONSE
United States	What is the impact on pipeline, conversion, and scale?	CRM integration, deployment speed, revenue measurement
Israel	Is the system robust, differentiated, and exportable?	Applied architecture, security, interoperability, rapid adaptation
Europe	Who controls the data, actions, and compliance?	Isolation, governance, audit, oversight, and regional hosting
Asia	Does the service work in our languages and channels?	Messaging, voice, always-on availability, linguistic and cultural localization

Table 11: Regional logics to incorporate into an international agentic-AI strategy.

SACHA OMON AI POSITIONING

Designed and developed in Israel for international markets, SACHA OMON AI combines a U.S.-style product mindset - revenue, speed, integration - with the control architecture expected in Europe and a multilingual, multichannel experience suited to Asian markets.

2.8 Designing in Israel with an architecture built for international markets

Designing a solution for multiple markets means accounting from the outset for differences in languages, channels, rules, business systems, and governance requirements. The architecture must be able to evolve without rebuilding the system for every new country or channel.

Israel is a particularly suitable anchor for this approach: language diversity, a strong technology culture, proximity to major international ecosystems, and a natural export orientation. That reality encourages teams to think about interoperability, security, and multilingual capability from day one.

For SACHA OMON AI, Israel is the anchor market. The United States, Europe, and Asia are target markets whose expectations are reflected in the design: value measurement and commercial integration in the United States; governance and control in Europe; language, voice, and mobile-use diversity in Asia. The point is not to claim deployment everywhere today, but to build an architecture that can progressively adapt to these environments.

LAYER	INTERNATIONAL FOUNDATION	REGIONAL ADAPTATION
Knowledge	Validated, versioned, and governed sources	Language, terminology, local rules, and access rights
Intelligence	Agentic orchestration and multi-model selection	Model, latency, processing location, and usage policy
Channels	Web, WhatsApp, phone, email, and SMS	Dominant channels, consent, cadence, and conversational conventions
Business systems	CRM, calendar, catalog, support, and logging	Connectors, hosting, identity, and organizational processes
Evidence	Quality, sources, actions, handoffs, and outcomes	Revenue, service, compliance, or experience metrics

Table 12: An architecture designed for international markets, with explicit adaptations by target market and organization.

The challenge is simple: design a foundation today that can operate tomorrow in multiple markets without losing knowledge, context, governance, or service quality. What changes from country to country are the rules, dominant channels, and expectations; the architectural principle remains common.

DEPLOYMENT PRINCIPLE

One central nervous system for knowledge, context, and governance; models, connectors, channels, and rules adapted to each market.

2.9 Reusing proof from one market to accelerate the next

International deployment becomes faster when the evidence produced in an initial market is structured for reuse. This is not about copying a commercial result outside its context, but preserving the assets that make it reproducible: evaluation scenarios, benchmark questions, validated sources, action limits, latency measurements, human-handoff rules, observed incidents, and acceptance criteria.

This discipline makes it possible to expand languages, channels, products, and countries while maintaining a consistent evidence standard: difficult cases, corrections, human decisions, and outcomes become an operational memory that accelerates subsequent deployments.

TYPE OF EVIDENCE	QUESTION DEMONSTRATED	REUSABLE ASSET
Technical	Does the system retrieve, reason, and act reliably?	Test sets, traces, sources, latency thresholds, and error rates
Business	Does the service actually improve the journey and the outcome?	Qualification, conversion, resolution, cost, and handoff metrics
Governance	Do data, permissions, and actions remain under control?	Policies, roles, approvals, logs, and incident procedures
Adoption	Do teams and customers use the service with confidence?	Field feedback, escalation reasons, training, and satisfaction criteria

Table 13: International proof rests on transferable assets, not on a general promise.

03

Customer cases and quantified evidence

Announcements are no longer enough. This chapter includes only what has been published, quantified, and dated.

Published results show what is already producing value; documented failures show what needs to be corrected. Both are presented with their source and the normal caveat that applies to vendor communications. A failure is not a final verdict on agentic AI: it must be viewed in the context of the maturity of the models, tools, and methods available when the project was built.

3.1 Published results

ORGANIZATION OR SCOPE	PUBLISHED RESULT	SOURCE
Salesforce: Agentforce customer base	More than \$100 million in cumulative annualized savings and a reported 34% productivity gain among customers; more than 18,000 customers in 124 countries.	Salesforce, page Agentforce Metrics
Salesforce: FY2026	Agentforce ARR of \$800 million, up 169% year over year; 2.4 billion agentic work units delivered across Agentforce and Slack.	Salesforce FY2026 results, February 2026
Salesforce: help site	More than 75% of inquiries resolved by the agent on a site receiving more than 60 million visits annually.	Salesforce case study
Nexo	62% of cases resolved by the agent after a phase focused on technical-debt remediation and data cleanup.	Salesforce case study
Safari365	15% efficiency target exceeded, with results above 30%; service delivered 24/7 across multiple time zones.	Salesforce case study
Wiley	Approximately 50% reduction in ramp-up time for advisors assisted by the agent.	2026 deployment benchmarks
PGA Tour (AWS)	Tournament editorial coverage produced ten times faster through a multi-agent system.	AWS, June 2026

ORGANIZATION OR SCOPE	PUBLISHED RESULT	SOURCE
Amazon Bedrock AgentCore	Tasks executed by agents increased fifteenfold in six months; SDK downloaded more than 2 million times in five months.	AWS, 2026
Demandbase (AI SDR agent for inbound)	Pipeline from the channel doubled, meetings with target accounts doubled, more than 100 SDR hours saved per month, approximately \$80,000 in savings.	Qualified / SaaStr case study
CX agent vendor (Sierra)	Estimated annualized revenue of \$200 million in May 2026, versus approximately \$130 million at the end of 2025: growth driven by resolution-based pricing.	Sacra, May 2026
CX agent vendor (Decagon)	More than 100 new enterprise customers in 2025; estimated annualized revenue of \$35 million in October 2025.	Sacra
Agent-executed financial reporting (multi-agent)	Production time reduced from 15 days to 35 minutes; cost per report reduced from approximately \$2,200 to \$9; average errors reduced from 3 to 0.3 per report.	Vendor / integrator case study

Table 14: Published results. Figures come from vendors or their customers and should be read with the normal caution applicable to commercial communications.

3.2 Academic evidence

- Finance. IIT Madras's multi-query agentic RAG pipeline improves accuracy by 11.2% and reduces hallucinations by 15% on standard financial question-answering datasets.
- Education. The MARAUS multi-agent system, evaluated on more than 6,000 real interactions, achieves 92% average accuracy, reduces the hallucination rate from 15% to 1.45%, and responds in under four seconds, at a two-week operating cost of less than \$12.
- Science. A hybrid agentic framework applied to scientific literature shows that hybrid retrieval provides the strongest foundation, with agentic reasoning adding controlled evidence expansion and grounded synthesis.
- Public sector. A statement-level faithfulness evaluation using multiple judge models highlights the contribution of tool-level attribution in multi-tool architectures to the interpretability and traceability of generated content.

Sources: arXiv 2509.16369; arXiv 2507.11272; arXiv 2606.10381; MDPI, Big Data and Cognitive Computing, 9(12), 309.

3.3 Documented failures: dated lessons, not verdicts

Three sets of public facts deserve to be presented because they define the conditions for success more precisely than success stories alone.

CASE	PUBLIC FACTS	LESSON
Klarna	The company announced that its AI assistant was doing the work of 700 customer-service agents, then brought human teams back after publicly acknowledging degraded quality and rising complaints. The system now operates as a hybrid model.	Fully agentic customer contact without calibrated human escalation damages the brand faster than it reduces costs.
11x	A TechCrunch investigation in March 2025 reported the use of logos from non-customers and overstated annualized revenue; customers reported performance below that of human teams.	A pilot on real-world data is worth more than a demo.
AI SDR market	An analysis of 100,000 matched emails published in April 2026 by Digital Applied measured a 4.1% response rate for AI-generated emails versus 5.2% for human-written emails, and a 0.7% meeting-booked rate versus 1.1%. The spam-report rate was 8% for AI versus 3% for humans.	Performance comes from qualification and grounding, not from send volume.
Customer support: claimed versus observed gap	A market audit published on August 15, 2026 reports vendor-claimed resolution rates of 65% to 86%, while published production data clusters between 40% and 70%. The gap is driven in part by different definitions of "resolution," request complexity, and knowledge-base quality.	Knowledge-base quality drives outcomes more than model choice.
Gartner forecast	More than 40% of agentic projects are at risk of cancellation by the end of 2027 because of poorly defined value, uncontrolled costs, and insufficient controls.	Value framing and governance are the first line of the project plan.

Table 15: Five lessons from publicly documented failures and performance gaps.

What these failures actually tell us

These cases remain useful, but they should not be read as permanent limits of agentic AI. A project launched two years ago was not built with the same models, platforms, integration capabilities, or usage habits available in 2026. Three developments change how we should read past failures.

1. Tools evolve. Building blocks that once had to be custom-built - memory, identity, retrieval, observability, connectors, controls - are progressively becoming platform components. The work shifts toward architecture, data, rules, and quality.
2. Usage and mindsets evolve. Customers, employees, and executives are becoming accustomed to asking AI for information or action. Adoption is not fixed: it increases with familiarity, trust, and service quality.

3. Development itself evolves. Programming, testing, documentation, and some integrations are increasingly AI-assisted. A project that is too long and too rigid can spend time building today what a platform may provide much more simply tomorrow.

FOUNDER'S CONVICTION - GLOBAL VISION, STEP-BY-STEP DEPLOYMENT

This is the method I propose: choose a real need, build a first capability, put it in front of real users, measure, correct, incorporate the advances that have appeared in the meantime, then expand. It is not a universal rule; it is my conviction for avoiding a frozen project while the technology keeps changing.

PART II

HOW DOES IT WORK?

A. Understanding the building blocks - LLMs, agentic AI, RAG, and why their combination changes everything.

04

Understanding LLMs: the AI you know - and its limits in the enterprise

It knows what is outside your company. It does not know what is inside.

This chapter begins with an observation you can test yourself in thirty seconds: the AI you already use answers very well when the information is public, and starts making mistakes when the question concerns the inside of your organization. We first describe that limitation and what it costs, then why it exists, and finally what must be added to overcome it.

4.1 An LLM is the AI you already know

Behind the acronym LLM - Large Language Model - is simply the language engine that powers the AI tools you use every day: ChatGPT, Claude, Gemini, Copilot. When an executive says, "I use AI," they are typically using an LLM.

This engine does two things, and it does them remarkably well. It understands a request stated in ordinary language, without keywords, forms, or special syntax. And it produces a clear answer adapted to the tone and context of the question. It also summarizes, translates, writes, reasons through steps, and produces code. These are real capabilities, and they explain the massive adoption of these tools.

A useful distinction for the rest of the White Paper: the LLM is the engine, not the product. ChatGPT is the engine plus an interface, memory, tools, and rules built around it. An enterprise agent follows exactly the same logic: it is not a model; it is a model plus what has been built around it. This entire White Paper is about what must be built around the model.

4.2 Outside vs. inside: the fundamental limitation

An LLM learned from public text. It therefore knows the outside of your company very well: what your website says, your brochures, press releases, articles that mention you, and customer reviews. Asked about that, it can answer accurately and quickly.

It does not know the inside. It does not know who does what in your organization, how a process really works, which terms you grant to which customer, what is in stock this morning, the status of a case, or what was promised in the last conversation. That information is not publicly written anywhere - it is private, live, and changes every day.

WHAT AI KNOWS ABOUT YOUR COMPANY	WHAT IT CANNOT KNOW
What the company sells, as presented on its website	The actual terms negotiated with each customer
Public messaging: brochures, press releases, media coverage	Internal procedures and the rules that are authoritative
Reputation: customer reviews, forums, rankings	Who does what, who approves what, who must be notified
The industry, competitors, and business vocabulary	Actual prices, inventory, and today's lead times
General knowledge up to the date of its training	A customer's history and the status of an active case

Table 16: The boundary is not technical; it is informational. The outside is public. The inside is private, organization-specific, and constantly changing.

Yet commercial decisions are made on the inside. An answer that knows only the outside is a brochure answer.

That is the technical reason for everything that follows. A company cannot act on a request if the system responding to it does not know its rules, pricing, procedures, and case status. Action begins with knowledge.

TAKE THE TEST - THIRTY SECONDS

Open ChatGPT or Claude and ask three questions about your own company.

1. What do you sell? - The answer is good: the information is on your website.
2. What is your refund procedure after thirty days? - The model invents a plausible answer, or tells you it cannot know.
3. What is the status of my customer Dupont's case? - Impossible.

You have just measured, using your own situation, where the outside ends and the inside begins. This entire White Paper is about the second part.

4.3 Six limitations of AI used on its own inside the enterprise

This boundary between outside and inside creates six concrete limitations. None is solved simply by switching models or choosing a larger model: each requires adding something around the engine. The third column shows what is needed and which chapter addresses it.

THE LIMITATION	WHAT IT MEANS IN PRACTICE	WHAT IS NEEDED TO OVERCOME IT
1. It does not know the inside of your company	Generic answers that are unusable with a customer; or pure invention when the question concerns an internal rule.	A private, governed knowledge base queried before every answer - RAG, Chapter 6
2. It does not know the present either	Its knowledge is frozen at the date of training: pricing, catalog, staffing, case law, and inventory can all be out of date or unknown.	Dated, versioned, and continuously updated knowledge - RAG, Chapter 6
3. It does not connect itself	The model does not read your CRM, open your calendar, or access your documents until it has been authorized and connected to do so.	Integration work: data, access rights, channels, oversight - Chapters 9 and 10
4. It does nothing	It generates text. It does not book a time slot, write to the CRM, follow up, or hand off a case.	A controlled ability to act - agentic AI, Chapter 5
5. Without a defined use case, it produces no measurable outcome	Everyone uses it on their own. Productivity remains individual; the company's process does not change.	A scoped and measured use case - the method, Part III
6. What you give it may leave the company	An employee who uploads a forecast, cash-flow plan, customer file, or contract into a consumer AI service is entrusting that information to a service the organization does not control.	A dedicated environment where data remains within the company's perimeter - Chapters 7 and 10

Table 17: Six limitations, six responses. None depends on model size.

The fifth is the most costly and the least visible. MIT Project NANDA's report estimates that, after \$30 billion to \$40 billion invested, 95% of the integrated pilots studied had produced no measurable P&L impact. This is not a model problem; it is a problem of undefined use and disconnected systems.

Primary source and status: preliminary findings, Project NANDA, July 2025, "The GenAI Divide: State of AI in Business 2025." Detailed analysis in Chapter 1.

The sixth limitation deserves more attention: it is the most widespread and the least visible

This limitation is organizational, not technical. Everywhere, employees are using AI on their own initiative, through personal accounts, usually because they are trying to do their jobs well: preparing a memo, summarizing a contract, building a forecast. This happens in private companies and in the public sector, including government agencies and ministries. The intent is good. The problem is that usage conditions differ depending on the account, and almost no one reads them.

WHAT THE CONTRACT SAYS - THAT ALMOST NO ONE READS

On consumer plans, exchanges may be used to improve models. At OpenAI, this is the default behavior, with an opt-out available in settings. At Google, this applies as long as activity history remains enabled - the initial setting. At Anthropic, since August 2025, users must make an explicit choice when creating an account.

On professional offerings - API, Team, Business, Enterprise - all three providers contractually exclude customer content from model training.

The difference therefore lies neither in the technology nor in the model: it lies in the contract purchased and the account being used. The same model, two different regimes.

Available measurements show the scale of the phenomenon. Across enterprise data sent to AI tools, the share classified as sensitive rose from 10.7% to 27.4%, then to 34.8% over three years. An analysis of 22.4 million workplace prompts identified 579,113 cases containing sensitive data. These are not leaks in the classic sense; they are voluntary transmissions made by well-intentioned employees.

Direct sources: Cyberhaven Labs, AI Adoption and Risk Report 2025 · Harmonic Security, analysis of 22.4 million workplace prompts, 2025 · OpenAI, Google, and Anthropic terms of use and privacy policies, reviewed August 19, 2026.

You do not build a sales forecast or cash-flow plan in a consumer AI environment if you cannot control what happens to the data you provide.

The reverse is also true, and it is the more useful argument for an executive. In a dedicated enterprise environment, security becomes manageable again: data remains within an identified perimeter, access is assigned, and usage is logged. From there, the tool can be opened to employees under controlled conditions. Each user can then draw on two bodies of knowledge instead of one: the model's general knowledge and the company's own knowledge, delivered through RAG. That is precisely the effect described in Chapter 7.

4.4 Why these limitations exist: how AI reads and how it answers

To understand why AI can invent an answer instead of admitting it does not know, you need to understand what happens between your question and its response. Two mechanisms explain most of it.

First mechanism: it does not read words; it reads numbers

A neural network does not manipulate letters or words directly; it manipulates numbers. Before processing begins, a component called a tokenizer breaks text into fragments - tokens - and assigns each one a numeric identifier. A token is neither a letter nor exactly a word; it is a frequent fragment of language. Common words may form one token; rare, long, or misspelled words are often split into several pieces.

WHAT YOU TYPE	WHAT THE MODEL SEES	NUMBER OF TOKENS
hello	a common word, kept whole	approximately 1 token
anticonstitutionnellement	a rare word, split into pieces	approximately 6 to 8 tokens
One page of standard text	approximately 500 words	approximately 650 to 800 tokens
A 40-page contract	the entire document sent to the model	approximately 30,000 tokens

Table 18: Illustrative orders of magnitude. Exact tokenization depends on each model's tokenizer and the language used.

A consequence that multilingual companies often overlook: for equivalent content, French generally uses more tokens than English. The same conversation therefore does not necessarily cost the same amount across languages.

Second mechanism: it predicts the most likely continuation

This entire industry rests on an architecture described on June 12, 2017 by a Google team in a paper titled "Attention Is All You Need": the Transformer. Its central idea can be summarized simply. Instead of reading text strictly word by word in sequence, the model looks across the elements of a sequence and calculates which ones matter most relative to the others. That is the attention mechanism, and it made today's language quality possible.

But at its core, what the model produces is still a prediction: the most likely next fragment given everything that came before. It has no native concept of true or false, only of what is statistically plausible. This is where the rest follows: when it has no information about your refund procedure, it does not necessarily stop - it produces the most plausible continuation. That is what we call a hallucination, and it is not simply a manufacturing defect; it follows directly from how the system is built.

A model does not "lie" in the human sense; it completes. Until you provide the inside of the enterprise, it may guess it.

Original document: "Attention Is All You Need," arXiv, June 12, 2017, Google Brain and Google Research. Full timeline: Appendix F.

4.5 What you pay for: the token is also the unit of sale

This mechanism has a direct economic consequence. Because everything the model needs must be provided in context during an exchange, and because that context is counted in tokens, the token has become the industry's commercial unit. Model providers typically meter input and output tokens separately rather than charging simply per question.

1

Input tokens

Everything the system sends to the model: the question, but also the conversation history, documents retrieved from your knowledge base, rules, and instructions. This is the largest and least visible cost component.

2

Output tokens

The text generated by the model. Usually priced higher than input tokens, but much lower in volume.

Two mechanisms can reduce this cost. Context caching avoids paying full price repeatedly for content that has already been transmitted, which matters for an agent that uses the same rules across interactions. Batch processing, when an immediate response is not required, is often priced at a discount. Terms vary by provider and change frequently.

In an agentic system, a single customer request can trigger several model calls: understand, retrieve, verify, formulate, act. Cost therefore depends not on the length of the answer displayed on screen, but on the number of steps and the amount of context used at each one. Token management becomes an architectural decision as much as a budget line.

The token is only one part of the total bill. The full cost of an agentic task - retrieval, tool calls, voice, storage, operations, and supervision - is detailed in Section 8.10.

4.6 The solution comes down to three building blocks

The six limitations described above reduce to three missing capabilities. The model lacks knowledge of the inside; it lacks the ability to act; and the company lacks control over what leaves its environment. Three responses correspond to those gaps.

WHAT IS MISSING	THE ARCHITECTURE THAT ADDRESSES IT	WHAT IT PROVIDES
Knowledge of the inside	RAG - Retrieval-Augmented Generation, Chapter 6	A private, dated, and governed knowledge base queried before every response: the model answers from your documents, not from its memory
The ability to act	Agentic AI, Chapter 5	An objective, a plan, authorized tools, observation of the outcome, and course correction until the task is completed
Control over what leaves the company	A dedicated and governed environment, Chapters 7 and 10	Data remains within an identified perimeter, access is assigned, and usage is logged: the tool can then be made available to all employees

Table 19: Three gaps, three responses. Chapter 7 shows why their combination produces more than the sum of the parts.

The next chapter presents the first of these architectures, which gives the model the ability to act, before Chapter 6 details the second, which gives it access to your enterprise knowledge.

05

Understanding agentic AI

Agentic AI marks the shift from intelligence that answers to intelligence that acts.

Unlike a model that simply produces text in response to an instruction, an AI agent receives an objective, builds a plan, mobilizes authorized resources, executes actions, and adjusts its path based on observed results. It combines reasoning, memory, tools, and bounded autonomy to produce a verifiable outcome. This chapter lays the foundations of that operating model.

FOUR RECOGNIZABLE DIMENSIONS

Intentionality - it pursues an objective defined by the user or by the system.

Autonomy - it chooses which actions to take and corrects itself when necessary.

Tool use - it uses external tools: document search, APIs, databases, calendars.

Continuous learning - it uses memory and feedback from experience to improve.

An AI agent does not merely understand. It acts. It adapts. It gets the job done.

SIMPLE DEFINITION

An agentic system receives an objective, understands the situation, prepares a plan, selects tools, observes what happens, corrects its path, and stops when the expected result is achieved, impossible, or must be handed to a human.

5.1 What "agentic" actually means

In a conventional interaction, the user gives an instruction and the software returns an output. In an agentic interaction, the user or process specifies a state to achieve: qualify a request, resolve an incident, prepare a proposal, verify eligibility, find an appointment slot, or update a case. The system must then choose and sequence multiple steps. The difference is less visible in the wording of the answer than in the responsibility delegated to the system.

The language model remains the engine for understanding and reasoning, but it is only one component. Around it sit external memory, rules, connectors, tools, a knowledge base, authorization controls, logging, and an escalation mechanism. That complete system is the agent. Calling an LLM an agent would be like calling an engine a car: it provides the central power, but not the steering, brakes, body, or traffic rules.

SYSTEM	WHAT IT RECEIVES	WHAT IT PRODUCES	RESPONSIBILITY
Rule-based chatbot	A predefined intent	An answer or a menu	Follow a closed scenario
LLM assistant	An instruction	A text, summary, or proposal	Help the human decide
Deterministic workflow	A defined event	A fixed sequence	Execute the predefined workflow
Agentic system	An objective and constraints	An outcome, an action, and a trace	Choose and adapt the path within a bounded scope

Table 20: Four families of systems that are often confused. Agentic systems are distinguished by an adaptive path, not by an absence of rules.

The market often calls a "deterministic workflow" automation: an architecture that executes a path fixed in advance, useful for a stable and repeatable process. An agentic system, by contrast, can choose its path at each interaction and adjust it based on what it observes. That capacity for choice is what distinguishes the categories, not the sophistication of the model used.

5.2 The agentic loop, step by step

DIAGRAM 3

The agentic loop

A business objective managed through to a verifiable outcome.

SCOPED AI AGENT	
1	OBJECTIVE Define the objective.
2	UNDERSTAND Interpret the request.
3	PLAN Choose the actions.
4	USE TOOLS Execute with the right levers.
5	OBSERVE Check the outcome.
6	ADAPT Correct and improve.

Stay focused · Iterate · Learn · Human oversight.

Figure 3: An agentic loop turns a business objective into a sequence of observable, controlled steps.

The objective provides direction. It describes the expected state and the limits to respect. "Help this prospect choose the right offer and secure agreement for an appointment" is agentic; "write an email" is an isolated task.

Understanding builds the situation. The system brings together the current message, contact identity, authorized history, channel, language, and applicable rules. It distinguishes what is known, what is missing, and what must be verified.

Planning chooses a path. Depending on the situation, the agent may answer directly, ask for clarification, retrieve information, call a tool, propose an action, or hand off to a human. The plan does not need to be long; it should be proportional to the task.

Tools produce real-world effects. Document retrieval, a CRM, a calendar, an inventory API, a payment solution, or a ticketing system become authorized functions. The agent does not merely simulate action: it executes it or prepares it for approval.

Observation checks the effect. After each call, the system verifies the result: is the appointment slot still available? Does the source actually answer the question? Was the record created? Failure becomes an input for control.

Adaptation prevents stubborn behavior. The agent can reformulate a search, choose another tool, reduce its ambition, abstain, or escalate. This ability to correct course is one of the most important traits of a production agentic system.

5.3 Answering is not yet acting

DIAGRAM 4

From response to execution

The benchmark is not how impressive the AI sounds; it is the task outcome.



| Cognitive task -> Traceable outcome.

Figure 4: The operational boundary lies between producing an answer and carrying a task through to a traceable outcome.

An answer can already be extremely useful: summarize a contract, draft a message, explain a procedure. Agentic capability begins when the system carries part of the operational continuity. It remembers that the

person already sent a document, checks what is missing, consults an up-to-date rule, proposes the next step, obtains approval, and then records the result in the system of record.

This distinction avoids two mistakes. The first is calling every assistant with a button or connector an "agent." The second is assuming an agent must be fully autonomous. In production, useful autonomy is almost always bounded: broader for reversible and well-documented actions, narrower for financial, legal, or human commitments.

5.4 Autonomy is a setting, not an ideology

Agentic AI does not require removing humans from the loop. It lets the organization decide precisely where humans intervene. An organization can require systematic approval at launch, then move to exception-based control, and later authorize certain low-risk actions once evaluations show stable quality. Autonomy is earned through evidence and withdrawn when the context changes.

DECISION TYPE	AGENT ROLE	RECOMMENDED HUMAN ROLE
Sourced factual information	Retrieve, synthesize, cite	Periodic source audit
Reversible action	Execute within predefined limits	Exception-based oversight
Commercial engagement	Prepare, calculate, propose	Approve exceptions and negotiations
Sensitive or ambiguous case	Detect risk, gather context	Decide and take over the relationship
Quality incident	Abstain, log, hand off	Correct the rule, data, or process

Table 21: Human-AI complementarity is designed decision by decision.

5.5 What agentic AI must not become

- A synonym for uncontrolled mechanical execution. A production agent operates within an explicit scope and knows when to hand off.
- A spectacular demo without metrics. Success is measured through service quality, task completion, satisfaction, traceability, and business impact.
- An omniscient model. The agent does not know proprietary knowledge until it is supplied through a governed knowledge layer.
- A black box without controls. Tool calls, sources, actions, permissions, and escalations must be logged; the model's private chain of thought does not need to be exposed.
- A uniform substitute for teams. The best use cases cover periods of unavailability, prepare work, maintain context, and increase service capacity.

KEY POINT BEFORE RAG

Agentic AI provides the ability to pursue an objective. By itself, it does not provide knowledge of the company. The next question therefore becomes unavoidable: what information should the agent rely on to make decisions? That is precisely the role of RAG.

An agentic agent is a system that receives an objective rather than a single instruction, develops a plan, selects and calls tools, observes results, corrects its path, and carries the task to a verifiable terminal state. The language model is the reasoning engine; the agent is the complete system built around it.

5.6 Six components of a production agent

COMPONENT	ROLE	COMMERCIAL ILLUSTRATION
Objective	The target state, expressed as a business outcome rather than a task.	Qualify an inbound request and book a qualified meeting.
Planning	Break the objective into ordered steps that can be revised during execution.	Identify the need, verify eligibility, propose a time slot, confirm.
Tools	Callable functions: document search, APIs, CRM, calendar, payments, telephony.	Read the calendar, create an opportunity, send a confirmation.
Memory	Short-term memory for the conversation, long-term memory for contact history.	Resume an exchange interrupted the day before on another channel.
Reasoning	Self-critique: is the response grounded, has the objective been achieved?	Detect missing information before proposing a time slot.
Guardrails	Deterministic policies, action scope, human escalation.	Immediate handoff to an advisor above a defined threshold.

Table 22: Anatomy of an agent. More of the value is concentrated in orchestration than in the underlying model.

5.7 Levels of autonomy in production

This scale does not measure model intelligence. It describes the operational responsibility delegated to the system and, in mirror image, the role retained by humans. Moving up one level only makes sense once scope, rights, evaluations, and human handoff are already under control.

DIAGRAM 5

Autonomy is a setting, not the absence of control

Each level has an explicit agent scope and an explicit human responsibility.

LEVEL	WHAT THE SYSTEM CAN DO	HUMAN ROLE
L0 • Assistant	Answers, summarizes, and produces content.	The human decides and executes every action.
L1 • Copilot	Proposes an action ready for approval.	The human approves every action before execution.
L2 • Supervised agent	Executes reversible actions within defined limits.	The human oversees by exception and handles alerts.
L3 • Bounded autonomous agent	Runs an end-to-end process within an authorized scope.	The human audits, sets policies, and takes over sensitive cases.
L4 • Multi-agent system	Coordinates multiple specialized agents and their tools.	The human governs identities, costs, permissions, and outcomes.

Increasing operational autonomy -> Oversight: from direct execution to governance.

Figure 5: Five levels of autonomy. The agent's scope expands while the human role evolves from direct execution toward governance.

LEVEL	DESCRIPTION	HUMAN OVERSIGHT
L0 • Assistant	Answers the request, produces text.	Full: the human executes.
L1 • Copilot	Proposes an action ready for approval.	Systematic approval.
L2 • Supervised agent	Executes reversible actions, escalates the rest.	Exception-based oversight.
L3 • Bounded autonomous agent	Runs an end-to-end process within a defined scope.	Post-hoc audit, alert thresholds.
L4 • Multi-agent system	Multiple specialized agents coordinated by an orchestrator.	Governance, observability, policies.

Table 23: Autonomy scale. The vast majority of profitable 2026 deployments sit at levels N2 and N3.

Field data supports this caution. McKinsey finds that high-performing organizations are much more likely to define explicit human-validation processes: 65% among high performers versus 23% among others. Autonomy is earned in stages and backed by measurement.

Source: McKinsey, State of AI, November 2025; CX Today analysis, February 2026.

5.8 The standard that unlocked the ecosystem: MCP

Connecting an agent to enterprise tools was long a custom integration project repeated for every model-tool pair. The Model Context Protocol, released as open source by Anthropic in November 2024, standardizes that connection: each application implements the protocol once, and each tool exposes it once.

- Adopted by competing players in less than eighteen months: OpenAI, Google, Microsoft, AWS, GitHub, and many other vendors.
- Governance of the protocol was transferred to the Agentic AI Foundation, a directed fund under the Linux Foundation, on December 9, 2025, with Anthropic, OpenAI, and Block as co-founders, and AWS, Google, Microsoft, Cloudflare, and Bloomberg as platinum members.
- Ecosystem scale: the official Registry listed more than 9,600 referenced servers by the end of May 2026; Anthropic cited more than 10,000 active public servers in its December 2025 ecosystem update.
- Enterprise adoption: a 2026 industry survey found that 41% of surveyed software organizations were running MCP servers in production, either on a limited or extensive basis.

Sources: Wikipedia, Model Context Protocol; Anthropic, December 2025 ecosystem update; Digital Applied, Registry API snapshot of May 24, 2026; Stacklok, 2026 software report.

WHY THIS MATTERS TO EXECUTIVE LEADERSHIP

An agent built on an open standard governed by a neutral foundation reduces vendor lock-in risk. This becomes a legitimate evaluation criterion in any agentic AI procurement process.

06

Understanding RAG: Retrieval- Augmented Generation

The model stops answering from memory. It answers from evidence.

RAG (Retrieval-Augmented Generation) combines the language-generation power of language models with the organization's validated knowledge sources. The system first retrieves relevant passages from a governed repository, then provides them to the model together with the question. The result: answers that are more accurate, current, and traceable.

RAG IN BRIEF - AUGMENTED ARCHITECTURE, NOT MAGIC

RAG does not replace the language model; it augments it. Before generating an answer, the system searches a document repository for relevant information and injects it into the generation context. This architecture improves factual accuracy, reduces hallucinations, and makes company-specific knowledge usable: internal policies, procedures, reference materials, FAQs, and business databases.

6.1 RAG defined: Retrieval-Augmented Generation

RAG connects the model to the organization's validated knowledge. Instead of asking the LLM to answer only from what it learned during training, the system first retrieves relevant passages from a governed document repository, then supplies them to the model with the question.

RAG is therefore neither an additional LLM nor an agent. It is a knowledge-access method. It provides facts, sources, and updateable context; on its own, it does not decide the path to follow, the tools to call, or when to hand the conversation to a human.

The previous chapter showed how an agent acts. But it must act on accurate information: that is the role of RAG. Query the knowledge base when the question is asked, then give the model the relevant passages so it can ground its answer in them. The model stops answering only from memory and starts answering from evidence.

6.2 Why not simply make the context window larger?

Language-model context windows have expanded dramatically, reaching hundreds of thousands of tokens at leading providers. This fuels a common objection to RAG itself: if an entire body of documentation can

now be inserted into a prompt, perhaps document grounding is unnecessary. Three operational realities still keep RAG at the center of production architectures.

- **Economics.** Inference cost rises with the number of tokens processed. Systematically injecting 300,000 tokens to answer a question that needs 2,000 can multiply cost dramatically on every production interaction.
- **Latency.** A conversational agent that takes ten seconds to answer risks losing the user. Targeted retrieval keeps response time compatible with a live conversation.
- **Precision.** Retrieval quality degrades for passages located in the middle of a very long context. Preselecting the most relevant passages improves the answer's fidelity to the source.

Market evidence supports this view: major platforms continued to invest heavily in managed document retrieval in 2026, even as their context windows reached unprecedented sizes. At its June 2026 New York Summit, AWS launched fully managed Knowledge Bases featuring an "Agentic Retriever" designed for complex, multi-step queries.

Source: AWS News Blog, AWS Summit New York announcements, June 2026.

6.3 The pipeline, step by step

DIAGRAM 6

The RAG pipeline

Nine steps: five offline to build the knowledge layer, four online to answer.

OFFLINE PHASE - STEPS 1 TO 5: BUILD THE KNOWLEDGE BASE

1	INGESTION Collect sources: documents, website, CRM, catalog, tickets, emails.
2	PARSING Structured content extraction: tables, PDFs, images, formatting.
3	CHUNKING Segment content into coherent, contextualized chunks.
4	VECTORIZATION Convert chunks into numerical representations.
5	INDEXING Store in a vector database, paired with a lexical index.

KNOWLEDGE BASE - RAG + MEMORY AND CONTEXT

ONLINE PHASE - STEPS 6 TO 9: ANSWER A QUESTION, WITH SOURCES

6	RETRIEVAL Retrieve relevant passages for the question being asked.
7	RERANKING Fine reranking of candidates by a specialized model.
8	GROUNDING GENERATION Generate an answer using only the supplied passages, with citations.
9	EVALUATION Measure faithfulness, relevance, and coverage.

Continuous improvement loop: evaluation feeds back into the offline steps.

Figure 6: The RAG pipeline - an offline indexing phase that feeds a knowledge base, followed by a response phase triggered by each question. Details of each step appear in Table 24 on the next page.

STEP	FUNCTION	KEY DECISIONS
1. Ingestion	Collect sources: documents, website, CRM, catalog, tickets, emails, business databases.	Scope, freshness, access rights, synchronization frequency.
2. Parsing	Structured content extraction: tables, PDFs, images, formatting.	Quality of extraction from tables and scanned documents.
3. Chunking	Segment content into coherent, contextualized chunks.	Chunk size, overlap, preservation of parent headings.
4. Vectorization	Convert chunks into numerical representations (embeddings).	Embedding-model choice, dimensionality, language, reindexing cost.
5. Indexing	Store in a vector database, often paired with a lexical index.	Per-customer isolation, metadata, security filters.
6. Retrieval	Retrieve the passages most relevant to the question.	Hybrid search (semantic and keyword), number of passages.
7. Reranking	Fine reranking of candidates by a specialized model.	Higher precision at the cost of additional latency.
8. Grounded generation	Generate the answer using only the supplied passages, with citations.	Instruction to abstain when a source is missing.
9. Evaluation	Continuous measurement of faithfulness, relevance, and coverage.	Business test sets, model evaluators, human review.

Table 24: The nine steps of a production RAG pipeline.

6.4 What RAG delivers, in numbers

- Accuracy and faithfulness. Recent academic work measures clear gains. A team at IIT Madras showed that, on financial question-answering tasks, a pipeline combining RAG with agentic orchestration improved accuracy by 11.2% and reduced hallucinations by 15%.

- Major reduction in a closed-domain setting. A documented university deployment - the MARAUS system, with more than 6,000 real interactions - reports average accuracy of 92% and a hallucination rate reduced from 15% to 1.45%, with response times under four seconds.
- Traceability. Each claim can point back to an identifiable source. This property is essential for use in regulated sectors - healthcare, finance, legal services, and the public sector - which, according to industry analysts, are among the leading adopters of RAG.
- Making enterprise knowledge usable. More than 80% of enterprise data is unstructured; RAG is the mechanism that makes it queryable by AI.

Sources: Srinivasan et al., IIT Madras, "Enhancing Financial RAG with Agentic AI and Multi-HyDE," arXiv 2509.16369; "An Empirical Study of Multi-Agent RAG for Real-World University Admissions Counseling," arXiv 2507.11272; Grand View Research.

RAG IN ONE SENTENCE

RAG replaces "the model thinks it knows" with "the model knows where to look - and cites what it read."

6.5 The limits of RAG used on its own

A classic RAG pipeline follows a fixed sequence: one question, one retrieval, one answer. That linear model runs into four ceilings.

1. Compound questions. "What is the delivery time for product X to Spain, and is it still in stock?" requires two separate searches across two different sources.
2. Multi-hop reasoning. When an answer depends on a fact established by a first search, a second search must be informed by the first.
3. No self-critique. A linear pipeline can produce an answer even when the retrieved passages are irrelevant.
4. No action. RAG can answer accurately; booking an appointment, writing to the CRM, and following up remain outside its scope.

These four ceilings map directly to the agentic architecture described in the previous chapter. The next chapter turns that complementarity into a system.

6.6 The knowledge base: the enterprise's first durable AI asset

The most realistic entry point for enterprise AI is not always a productivity project. It can be simpler, more useful, and easier to align around: making existing knowledge accessible, governed, and usable. An organization already owns most of the raw material in its web pages, offers, procedures, catalogs, standard contracts, support answers, meeting notes, CRM, and the memory of its experts. The problem is fragmentation, obsolescence, and access rights.

Building a knowledge base for RAG does not mean dumping every file into a search engine. The organization must identify the authoritative source, the editorial owner, the validity date, the authorized audience, the update frequency, and what the system should do when information is missing. This work creates a knowledge layer that multiple systems and teams can reuse. Even if the first use case is sales, the resulting asset can later support customer service, onboarding, operations, training, or internal assistance.

BEFORE	TRANSFORMATION	RESULTING ASSET
Scattered documents	Inventory and classification	Known and attributed corpus
Conflicting versions	Authoritative source and validity dates	Consistent answer
Knowledge held by a few people	Formalization and validation	Transferable organizational memory
Undifferentiated access	Metadata and access control	Permission-aware retrieval
Occasional update	Synchronization and alerts	Living knowledge
Unmeasured answers	Question sets and evaluation	Observable quality

Table 25: A RAG project turns a stockpile of files into a governed knowledge layer.

This layer becomes an AI development foundation for the entire company because it separates knowledge from the channel. The same validated facts can power an agent on the website, a WhatsApp conversation, a phone call, an email exchange, an SMS notification, an internal assistant, or an after-sales journey. Wording and action rules change; the source of truth remains shared. The enterprise avoids creating five inconsistent assistants that separately learn five versions of the same offer.

ARCHITECTURE PRINCIPLE

Knowledge must be shareable without becoming undifferentiated: a common foundation, explicit access rights, views adapted to each function, and strict isolation of data specific to each customer or entity.

6.7 One knowledge layer across web, WhatsApp, phone, email, and SMS

RAG is not tied to an interface. It sits after the request is understood and before the response or action. On the web, the input is textual and can be enriched by the page being viewed. On WhatsApp, the conversation is asynchronous, persistent, and often made up of short messages, attachments, or voice notes. On the phone, speech must be transcribed, processed, and returned with very low latency. In all three cases, the agent can query the same knowledge base, apply the same rules, and preserve a unified relationship context.

CHANNEL	WHAT THE AGENT RECEIVES	REQUIRED ADAPTATION	HUMAN VALUE
Web	Text, page, campaign, browsing behavior	Immediate response, links, forms, and appointment booking	The sales rep receives an already contextualized request
WhatsApp	Messages, history, documents, audio	Asynchronous conversation, consent, resume without repetition	The advisor picks up the exact thread at the right moment
Phone	Voice, pace, interruptions, intent	Real-time transcription and summarization, short sentences, seamless escalation	The human steps in for emotion, complexity, or negotiation
Email	Long message, attachments, conversation thread	Structured response, citations, timing, and possible approval	The expert approves sensitive commitments
SMS	Short message and mobile identity	Concise response, links, and reminders	The appointment and continuity are secured

Table 26: The channel changes the form of the interaction, not the source of truth.

Multichannel consistency does not mean copying the same message everywhere. A sound architecture separates three layers: knowledge, which is shared and validated; reasoning, which determines intent and the next step; and expression, which is adapted to the channel. A phone response should be short and natural, while an email can be more detailed and documented. What must not vary are pricing terms, eligibility rules, promise limits, or customer history.

This model also makes internal adoption easier. The organization can start with one knowledge scope and one low-risk channel, measure quality, then expand gradually. Each extension reuses the document asset, evaluation sets, rules, and history already built. Deployment becomes cumulative rather than a succession of isolated prototypes.

6.8 Securing knowledge: a governed architecture, not simple copy-and-paste

RAG does not make information secure simply by indexing it. Security comes from the architecture around it. An enterprise knowledge base must preserve source access rules, separate each organization's environment, filter retrieval by identity and role, encrypt data at rest and in transit, protect connector secrets, limit retention, and log both consultations and actions. The agent should retrieve only what the person and the use case are authorized to know.

RISK	EXPECTED CONTROL IN AN AGENTIC RAG ARCHITECTURE
Data leakage between users or customers	Workspace isolation, role-based access control, and filters applied before retrieval
Sensitive document retrieved out of context	Sensitivity metadata, source-inherited permissions, and exclusion policies
Prompt injection or malicious document	Source validation, content analysis, separation between instructions and retrieved documents
Unauthorized business action	Least-privilege permissions, risk-based human approval, and an explicit allowlist of tools
Lack of evidence	Logs of consulted sources, answers, decisions, tools called, and human handoffs
Excessive retention	Defined retention periods, controlled deletion, encryption, and contractual choice of model provider

Table 27: RAG security depends on explicit controls at every layer.

The most common risk is not the name of a tool; it is using an assistant without an enterprise framework. Copying confidential information into a personal account, without an internal policy, settings control, rights management, or contractual data commitments, creates avoidable exposure. Terms differ by offering. OpenAI, for example, states that individual services may use content to improve models unless the user opts out, while ChatGPT Business, ChatGPT Enterprise, and the API do not use an organization's inputs and outputs for model training by default.

An architecture such as Sacha Engine must therefore select the model and connection mode according to sensitivity level, location, retention policy, and required contractual commitments. Dedicated RAG keeps knowledge inside a governed perimeter; the orchestrator enforces rights and rules; the LLM receives only the context needed to answer; sensitive actions require additional control. Knowledge protection becomes a verifiable system property rather than a general promise.

SECURITY PRINCIPLE

The right question is not only "which model are we using?" but "what data are we sending to it, for which user, with which permissions, for how long, and with what audit trail?"

References: NIST AI 600-1, OWASP LLM08:2025, and OpenAI Business Data Privacy; details in Appendix D.

07

Agentic AI + RAG + LLM: why the combination wins

Agentic AI decides and acts. RAG provides company knowledge. The LLM understands and formulates. Separately, none is enough.

Taken separately, the three building blocks each solve only part of the problem: an agent without governed knowledge acts on insufficient information; RAG without an agent retrieves information but never drives a journey; an LLM alone can formulate well, but from its training memory rather than your data. Combined, they form an operational system - informed, traceable, and controllable - whose assembly this chapter describes layer by layer.

Before we assemble them, remember this: these three building blocks are also the three questions to ask any AI-agent vendor. They require no technical expertise, and the answers will tell you in three minutes what kind of system you are dealing with.

THE QUESTION TO ASK	WHAT IT VERIFIES	THE ANSWER THAT SHOULD RAISE A RED FLAG
"Which model do you use, and can I switch?"	The LLM: the ability to understand and formulate - and your freedom not to depend on a single provider.	"It is our proprietary AI," without naming the model or providing for model portability.
"Where does the answer come from when the agent talks about my prices, offers, or customer cases?"	RAG: up-to-date company knowledge that is sourced and verifiable sentence by sentence.	"The model was trained on your data" or "it learns by itself from conversations."
"What can it do, and how far can it go without me?"	Agentic AI: the ability to act - propose a time slot, send a document, follow up, hand off - and the limits of that autonomy.	"It answers every question," with no list of actions and no human-handoff rule.

Table 28: Three questions, three building blocks. If one is missing, the agent will have a structural weakness: it will make things up, it will not know, or it will not act.

These three questions do not explain how the system is built. They tell you whether it has the three essential building blocks - in other words, whether it can function at all. Implementation is a separate issue: tools and integrations, memory, guardrails, evaluation, and human handoff. Those are validated in the project requirements document in Part III.

7.1 Integrating LLMs, RAG, and agentic AI into the enterprise

The teaching order should not be confused with the technical flow. To understand the system, we first distinguish agentic AI, then RAG, then the LLM. In operation, these building blocks interact in a loop: a request enters through a channel; the agent identifies the goal and the rules; RAG retrieves the authorized information; the LLM interprets and formulates; then the agent sends the response, calls an authorized tool, or hands off to a human.

STEP	PRIMARY BUILDING BLOCK	ROLE IN THE ENTERPRISE
01	Request and channels	Web, WhatsApp, phone, email, or SMS: the person expresses a need in their own language and at the moment it arises.
02	Agentic layer	It understands the objective, chooses the steps, applies the rules, selects authorized tools, and knows when to hand off to a human.
03	RAG and knowledge base	They retrieve relevant passages from the company's validated content, with access rights, versions, and sources.
04	LLM	It interprets the question and the retrieved information, then prepares a clear, contextualized response adapted to the language and channel.
05	Systems, human, and traceability	The system delivers the response, updates authorized tools, preserves context, and hands the request to the right team when human intervention creates value.

Table 29: The five steps connecting an inbound request to action, and the building block responsible for each one.

This separation avoids confusion: an LLM alone can formulate but does not necessarily know the enterprise; RAG alone can retrieve but does not drive a journey; an agentic layer without governed knowledge may act on insufficient information. Value appears when all three are connected to the organization's real teams, rules, tools, and channels.

PUBLIC ARCHITECTURE · SACHA OMON AI

In this architecture, Sacha Engine is SACHA OMON AI's proprietary orchestration layer. It connects channels, a selectable intelligence layer, the enterprise's governed knowledge, authorized actions, and human handoff. The public diagram below shows this division of responsibilities without disclosing the internal logic for routing, context construction, memory, Corrective RAG (CRAG), control, or execution.

DIAGRAM 7

SACHA OMON AI operating architecture

Sacha Engine connects channels, intelligence, knowledge, and human handoff.

CHANNELS	Web · WhatsApp · Phone · Email · SMS
SACHA ENGINE	Agentic AI agent · proprietary, secure, confidential orchestration
INTELLIGENCE AND KNOWLEDGE	Selectable LLM · dedicated RAG
RESPONSE AND HUMAN HANDOFF	Always-on service · context transmitted

Proprietary orchestration · secure · confidential.

Figure 7: Public architecture of SACHA OMON AI. Sacha Engine orchestrates the AI agent, intelligence layer, enterprise knowledge, and human handoff without exposing proprietary mechanisms.

7.2 RAG becomes agentic: the agent decides when to search and when to act

DIMENSION	CONVENTIONAL RAG	AGENTIC RAG
Trigger	Systematic search on every request.	Retrieval decided by the agent based on the actual need.
Query	The user's question, as stated.	Question decomposed and reformulated into several complementary queries.
Sources	A single index.	Routing to the right source: documents, CRM, catalog, graph database, web.
Iteration	A single pass.	Successive loops, with each result informing the next search.
Quality control	Direct generation.	Self-critique: assess whether the evidence is sufficient before answering.
Outcome	A written response.	A response plus an action executed in business systems.
Traceability	Citation of the passages used.	Full audit trail: plan, tool calls, sources, decision, action.

Table 30: From a linear pipeline to a reasoning-driven loop.

7.3 Why agentic RAG improves outcomes

1. Query decomposition. The agent breaks a compound question into independent subquestions, then recombines the results. Research from IIT Madras applies this principle through the generation of multiple non-equivalent queries and measures combined gains in accuracy and hallucination reduction.

2. Routing to the relevant source. A pricing question queries the catalog; a status question queries the management system; a method question queries the documentation. Routing prevents relevance from being diluted inside a single index.
3. Iterative multi-hop retrieval. The agent chains searches based on facts already established, a capability explicitly measured by multi-hop evaluation sets.
4. Hybrid retrieval and reranking. Combining lexical and semantic search, followed by reranking, is described in recent literature as the most robust retrieval foundation; agentic reasoning then supports controlled evidence expansion and synthesis.
5. Self-critique and abstention. The agent evaluates whether the evidence is sufficient and can declare that it cannot answer rather than improvise - a critical behavior for trust in commercial or regulated contexts.
6. Persistent memory. The agent retains the contact's history and incorporates it into later searches, directly addressing the weakness identified by MIT: generic tools often fail to retain context from one session to the next.

Sources: arXiv 2501.09136, "Agentic Retrieval-Augmented Generation: A Survey"; arXiv 2509.16369; arXiv 2606.10381; MIT NANDA, 2025.

7.4 What the research establishes - and where it adds nuance

The leading academic synthesis on agentic RAG provides two useful lessons for an investment decision.

- The gains are strongest in domains with structured knowledge and explicit constraints: healthcare, finance, and law. Open-ended tasks show more modest gains. In other words, domain modeling remains a core component of the system.
- The cost of additional reasoning must be measured. A comparative study published in 2026 found that decomposition delivers domain-dependent benefits and that reflection loops add significant latency for uneven quality gains. The design should therefore calibrate the level of iteration to the actual difficulty of the request.

Sources: arXiv 2501.09136; "Agent-Orchestrated Adaptive RAG," arXiv 2606.05658.

THE DESIGN PRINCIPLE TO REMEMBER

A useful agent applies proportional effort: direct response for simple questions, a single retrieval step for factual questions, and decomposition plus iteration for complex questions. Uniform sophistication increases latency and cost without improving perceived quality.

7.5 The reference architecture

SACHA OMON AI application: in this diagram, the "agent" layer corresponds to the SACHA OMON AI agent, orchestrated by Sacha Engine.

DIAGRAM 8

Five-layer reference architecture

From channels to governance, which spans the entire system.

1 · CHANNELS - CONVERSATION INPUT AND OUTPUT	Website · WhatsApp · Phone · Email · SMS
2 · ORCHESTRATION - REASONING, PLANNING, TOOL SELECTION, REFLECTION	Agent: agentic reasoning engine
3 · KNOWLEDGE AND MEMORY - WHAT THE AGENT KNOWS AND RETAINS	Knowledge (RAG): vector database · lexical index · graph · business connectors - Memory: session · contact profile · multichannel history
4 · TOOLS - ACTUAL EXECUTION OF ACTIONS	CRM · Calendar · Billing · Telephony · Payments
5 · GOVERNANCE - SECURITY, COMPLIANCE, AND QUALITY ACROSS ALL FOUR LAYERS	Logging · guardrails · continuous evaluation · human escalation

Figure 8: Reference architecture applied to the SACHA OMON AI agent. Layer 3 contains two components - knowledge (RAG) and memory - detailed separately in Table 31.

LAYER	FUNCTION	TYPICAL COMPONENTS
Channels	Conversation input and output.	Website, WhatsApp, telephony, email, SMS.
Orchestration	Reasoning, planning, tool selection, reflection.	Agentic engine, policies, state machine.
Knowledge	Factual grounding for every response.	Vector database, lexical index, graph, business connectors.
Memory	Conversational and relationship continuity.	Session memory, contact profile, multichannel history.
Tools	Actual execution of actions.	CRM, calendar, billing, telephony, email, payments.
Governance	Security, compliance, quality.	Logging, guardrails, continuous evaluation, human escalation.

Table 31: Reference architecture for a RAG-augmented agentic AI agent.

7.6 From the first message to a multichannel response: the real-world journey

The public diagram below shows what the user experiences without exposing Sacha Engine's proprietary mechanisms. A person starts a conversation on the web, WhatsApp, phone, email, or SMS. SACHA OMON AI mobilizes the required intelligence and the enterprise's governed knowledge to understand the request, provide a useful response, qualify the situation, and prepare the right next step on the appropriate channel.

The enterprise benefits from a continuous experience and a shared knowledge layer across channels. Internal orchestration, model routing, context construction, memory, control, CRAG, and tool-calling

mechanisms are part of Sacha Engine's proprietary know-how and are intentionally not detailed in this White Paper.

DIAGRAM 9

From the first message to a multichannel response

Every layer interacts with the others; humans remain at the center of judgment, approvals, and sensitive decisions.

HUMAN - JUDGMENT · NEGOTIATION · DECISION-MAKING · APPROVAL OF SENSITIVE CASES · GOVERNANCE AND ACCOUNTABILITY

1	PROSPECT OR CUSTOMER Web · WhatsApp · phone · email · SMS. Question, request, intent.
2	SACHA OMON AI Agentic AI agent · Sacha Engine. Understands, reasons, prepares the appropriate response and action.
3	ENTERPRISE KNOWLEDGE Dedicated and governed RAG. Facts, rules, context.
4	MULTICHANNEL EXPERIENCE Response · qualification · action.

Two-way exchanges: humans orchestrate and step in at the right moment. Sacha processes and connects interactions; humans receive, validate, arbitrate, and redirect whenever their judgment adds value.

Figure 9: From the first message to a multichannel response. Exchanges are bidirectional, and humans remain at the center of judgment, approvals, and governance.

7.7 The intelligence layer: where LLMs connect

In the SACHA OMON AI architecture, the LLM is a selectable intelligence layer. It understands, reasons, and formulates, but it is neither the knowledge base, nor the enterprise memory, nor the holder of permissions. Sacha Engine is the proprietary orchestration layer; dedicated RAG maintains the enterprise's governed knowledge; humans retain judgment, exception handling, and responsibility.

This separation is strategic: the enterprise is not locked into a single model, and its knowledge does not become the property of the LLM. The model receives only the context required for the task, within the framework defined by the architecture. Provider choice, reasoning level, latency, and data constraints can evolve without moving the central asset: the enterprise's knowledge and rules.

For the user, this complexity remains invisible. The user asks a question through a channel; SACHA OMON AI responds in the same context, continues the conversation, and hands off to a human when the situation requires it. The White Paper explains this division of responsibilities so the architecture can be understood without publishing Sacha Engine's proprietary routing, control, memory, or execution logic.

DIAGRAM 10

Four complementary roles, one human decision

How AI works with the enterprise.

1 · LLM Understands and formulates It interprets the request and creates a natural, useful response.	2 · RAG Provides company facts It retrieves validated knowledge: offers, rules, documents, and context.	3 · AGENTIC AI Guides the journey It chooses the next action, follows up, qualifies, and routes at the right moment.	4 · HUMAN Retains accountability The human exercises judgment, approves sensitive decisions, and has the final say.
--	---	--	---

AI becomes an operational capability of the enterprise. It responds where value is lost, while remaining orchestrated under human oversight.

Figure 10: Four complementary roles distributed across the six technical components detailed in Table 32 - the LLM understands and formulates; RAG provides enterprise facts; agentic AI drives the journey; humans retain responsibility and the final say.

LAYER	WHAT IT IS FOR	WHAT IT MUST NOT PRETEND TO BE
LLM	Understand, reason, plan, synthesize, and formulate.	The company's source of truth, memory, or permissions system.
Agentic orchestrator	Choose the model, route the request, apply policies, manage state, and determine the next step.	A simple prompt sent once.
Dedicated RAG	Retrieve authorized, current, relevant passages; preserve links to sources.	Model training or undifferentiated document storage.
Memory and context	Preserve identity, authorized history, case status, and the next action.	Implicit or uncontrolled memory held by the model provider.
Business tools	Create an appointment, read a status, enrich a CRM, open a ticket, or prepare a payment.	An action merely simulated in a text response.
Governance	Control identities, permissions, approvals, costs, evaluations, traces, and human handoffs.	A control layer added after deployment.

Table 32: Each layer has a distinct responsibility. Robustness comes from their coordination, not from the LLM in isolation.

KEY TAKEAWAY

The LLM is the engine for understanding and reasoning. RAG provides company facts. The orchestrator guides the journey. Memory preserves authorized context. Tools create the real-world effect. Governance controls the whole system.

7.8 A multi-model architecture instead of dependence on one provider

Sacha Engine can connect several model families: OpenAI GPT, Anthropic Claude, Google Gemini, xAI Grok, as well as open models such as Llama or Mistral when the context warrants it. The goal is not to accumulate providers, but to avoid forcing the same latency, cost, reasoning level, and data constraints onto every use case through a single model choice.

Multi-model capability allows the intelligence level to be adapted to the context without publishing the internal selection rules. A real-time interaction, a complex analysis, a multilingual task, or data subject to regional constraints do not necessarily require the same model. Sacha Engine applies the policy defined for the enterprise while preserving a provider-independent architecture.

This independence also protects the central asset: enterprise knowledge. Documents, metadata, permissions, histories, and evaluation sets should not be locked inside a model. They remain in the Sacha Engine layer and can be made available to the LLM best suited to the situation.

Sources: official documentation from OpenAI (Agents), Anthropic (Claude and MCP), Google Cloud (Gemini and Vertex AI Agent Builder), xAI (Grok and agentic tools), Meta (Llama), and Mistral AI.

B. FROM REQUEST TO COMMERCIAL ACTION

The architecture applied to the use case where value disappears fastest.

08

From request to immediate commercial action: the multichannel agentic AI SDR agent

The exact place where value is lost is between the paid click and the response delivered.

The opening pages of this White Paper established a behavioral shift: when a need appears, we are increasingly accustomed to getting an answer immediately. This chapter examines what that expectation changes in the sales relationship - capturing demand at the moment it exists, responding, qualifying it, documenting it, and then handing it to a human with the full context.

8.1 Demand no longer waits: real time becomes the new sales rule

In 1996, the Internet taught us to search for information. In 2026, AI is increasingly teaching us to state a need directly and receive information, a recommendation, a service, and sometimes even an action immediately. This shift in mindset is not confined to personal assistants; it is gradually changing what people expect from companies as well.

When a prospect clicks an ad, opens WhatsApp, lands on a website, or calls, the intent exists now. The prospect does not think in terms of sales-office hours or average callback time. The first requirement is therefore to start the relationship immediately: understand why the person is there, provide a useful first response, clarify the need, and prepare the next step.

The numbers show that sales has already begun this shift. According to Salesforce's 2026 State of Sales study of more than 4,000 sales professionals, 87% of sales organizations already use AI somewhere in the sales cycle - prospecting, forecasting, lead scoring, or email drafting. The question is no longer simply whether AI is entering sales, but how it connects to the exact moment demand appears.

IBM describes how an AI SDR agent can work: as soon as a prospect submits a form or starts a conversation, the agent can ask qualification questions, interpret the answers, and check them against criteria defined by the company without waiting for an employee to become available.

McKinsey adds the point that matters most for senior leadership: value does not come from adding an AI button to an existing journey, but from redesigning the workflow around the agent. Its research on agentic marketing workflows discusses as much as two-thirds of current marketing activities and a 10% to 30% revenue lift associated with hyperpersonalization.

Verizon illustrates this shift at scale. According to Reuters, by April 2025 an AI assistant deployed to 28,000 customer-service representatives had reduced call duration and freed up selling time, with sales up nearly 40% since deployment.

THE THROUGH-LINE OF CHAPTER 8

Breakthrough 1: the request expects an immediate response. Breakthrough 3: source, consent, and proof become operational data. The AI SDR agent sits exactly at the intersection of these two shifts.

Sources: Salesforce, State of Sales 2026, 7th edition, more than 4,000 respondents, February 2026; IBM Think, study on how AI SDR agents are reshaping sales, April 2026; McKinsey, "Reinventing marketing workflows with agentic AI," 2026; Reuters, April 2025.

APPLICATION: SACHA OMON AI

SACHA OMON AI applies this principle directly: its multichannel agentic AI SDR agent responds in under two seconds, 24/7, across web, WhatsApp, phone, email, and SMS, using a RAG knowledge base dedicated to each customer company.

8.2 Market proof: one hundred million prospects never called back - and Salesforce's acquisition of Qualified

An episode at the world's largest CRM vendor illustrates, better than any study, the scale of the problem an agentic AI SDR agent can address.

100 million

Leads left unanswered over roughly a quarter century, according to Marc Benioff
The Logan Bartlett Show, August 29, 2025

10,000

Leads an agentic representative can call back in one week
Example cited by TIME, August 28, 2025

68,000

Leads analyzed and re-engaged by Salesforce
Figure published as of October 1, 2025

Marc Benioff explained that Salesforce had accumulated roughly 100 million leads without follow-up because there was not enough human capacity. Salesforce later documented its use of Agentforce to analyze, reengage, and qualify those leads. As of October 1, 2025, the company reported 68,000 newly analyzed and reengaged leads, 800 meetings booked automatically, and dozens of closed deals.

WHAT THIS CASE SHOWS

A company can invest heavily in acquisition and still lose value after the click if the request is not handled. The issue is therefore not only generating more leads: it is being able to respond when intent exists, then handing to humans the cases that require judgment.

Editorial note: this public case is cited as a market reference. It does not constitute a partnership with, or endorsement of, SACHA OMON AI by Salesforce or Marc Benioff.

On August 29, 2025, Marc Benioff, co-founder, chair, and CEO of Salesforce, said during an interview on The Logan Bartlett Show that more than 100 million prospects who had contacted Salesforce over the previous 26 years had never been called back because the company lacked sufficient human capacity. Salesforce has since deployed its own agentic sales technology to engage more than 10,000 inbound prospects per week and turn those conversations into sales pipeline.

Four months later, on December 17, 2025, Salesforce announced a definitive agreement to acquire Qualified, a provider of agentic marketing and AI SDR solutions whose Piper product continuously qualifies and converts inbound website visitors. The acquisition closed on April 1, 2026. In its Form 10-Q for the quarter ended April 30, 2026, Salesforce reported that the fair value of the consideration transferred for Qualified was \$1.2 billion, primarily \$1.1 billion in cash.

The transaction is a market signal in its own right: technologies capable of engaging, qualifying, and converting inbound demand without breaking continuity are now valued at the billion-dollar level - including by the company that had just publicly demonstrated the internal cost of not having that capability.

Sources: Marc Benioff, The Logan Bartlett Show, August 29, 2025; Salesforce, December 17, 2025 announcement, updated April 1, 2026; Salesforce Form 10-Q, quarter ended April 30, 2026.

8.3 Markets are beginning to price in AI risk for contact centers

Financial markets have begun to price in a strategic risk for large contact-center operators: when AI agents can understand a request, consult a knowledge base, respond on the web or by phone, and hand context to a representative, value can no longer rest solely on the number of people available.

On February 28, 2024, Teleperformance shares fell 27% in a single session and reached a seven-year low, with Reuters explicitly linking the move to concerns about AI's ability to transform the call-center business. The figure on the next page compares the two-year stock performance of Teleperformance and three comparable U.S.-listed groups.

DIAGRAM 11

Stock market · customer-service operators: the two-year revaluation

Unadjusted closing prices · comparison from August 13, 2024 to August 13, 2026.

COMPANY	CLOSE ON AUGUST 13, 2024	LATEST CLOSE	CHANGE
Teleperformance — TEP · Euronext Paris · EUR	103,10 €	69,10 €	- 33,0 %
Concentrix — CNXC · Nasdaq · USD	64,53 \$	24,51 \$	- 62,0 %
TTEC Holdings — TTEC · Nasdaq · USD	4,32 \$	1,57 \$	- 63,7 %
TaskUs — TASK · Nasdaq · USD	13,40 \$	6,92 \$	- 48,4 %

Reading the chart: over two years, the four companies lost between 33.0% and 63.7% of their market value. Latest close as of August 13, 2026, and August 12 for Nasdaq-listed names: at the August 13 data pull the session was still in progress, so the last complete close was used to avoid intraday pricing.

Figure 11: Two-year stock-market performance of Teleperformance and three U.S.-listed contact-center groups. Sources: Yahoo Finance (TEP, CNXC, TTEC, TASK price histories) · Euronext Paris · TTEC Investor Relations and LSEG · TaskUs Investor Relations and LSEG · SACHA OMON AI calculations.

These prices do not prove that AI alone explains the decline in valuations. Organic growth, margins, debt, customer mix, and execution quality remain decisive. They do show, however, that investors are reassessing models in which service capacity depends primarily on the number of representatives available. Teleperformance is not standing still: the group, which says it operates in nearly 100 countries, 400 languages, and 170 markets with almost 490,000 employees, is developing its own TP.AI offering and an orchestration model combining AI and people.

THE SIGNAL TO REMEMBER

Human relationships are not disappearing. What is becoming fragile is the model in which value depends only on the number of available advisors. The future belongs to architectures that make knowledge immediately accessible, expand service capacity, and reserve for humans the situations where judgment creates the difference.

Source for the 2024 market signal: Reuters, February 28, 2024. Price sources and methodology: shown in the adjacent figure.

8.4 What humans cannot do at scale: answer everything, all the time

Productivity is not the first sales benefit of agentic AI. The first benefit is structural: an AI agent can do what no human organization can reproduce economically - handle every inbound request, at any hour, without a queue or a day off. That capability changes the nature of the customer experience far beyond a simple time saving.

The gap can first be measured in advertising budgets. EMARKETER forecasts about \$243 billion in global digital advertising revenue for Meta in 2026 and \$240 billion for Google; in the U.S. market alone, annual digital ad spending is estimated at \$413 billion. These analyst forecasts are not directly comparable with the advertising revenue reported by the two groups in their official accounts (page 4), whose scope is broader. What they consistently show, however, is that every dollar spent on these campaigns is intended to generate inbound demand.

Yet a significant share of that demand arrives outside business hours: HubSpot puts the share of phone inquiries received outside standard hours at 52%, a magnitude echoed by several 2026 industry studies for web forms and appointment requests.

A study of 22,500 U.S. auto dealerships conducted by Foureyes and reported by Ringlead in 2026 quantifies the cost of this gap. For a dealership spending an average of \$45,000 per month on advertising and generating about 150 inbound inquiries, 43% never receive an actual phone call - meaning nearly \$19,350 in monthly ad spend generates prospects who are never contacted by a human voice. The same work estimates roughly a \$95,400 monthly margin gap between average response performance and a response time under sixty seconds.

Sources: EMARKETER, digital advertising forecasts, April 2026; ResearchAndMarkets, United States Digital Ad Spend Market, February 2026; HubSpot; Foureyes and Ringlead, study of 22,500 U.S. auto dealerships, 2026.

This unhandled demand is not just an after-hours problem. Web forms impose delay and friction that today's buyer tolerates less and less; the previous generation of chatbots, built on fixed decision trees, failed on any question outside the expected script - exactly the limitation detailed in Chapters 5 and 6. An agentic AI SDR agent addresses both obstacles at once: it answers accurately from RAG-grounded knowledge, then routes the conversation according to its purpose - meeting with a decision-maker, connection to a sales representative, or direct handling of a customer-service request - instead of leaving the person without an answer, waiting for a callback that may never come.

WHAT AGENTIC AI CHANGES STRUCTURALLY

A human call center, even one operating 24/7, remains limited by the number of advisors available at any given moment and by the breadth of their expertise when handling objections. An agentic AI agent can handle simultaneous conversations without being constrained by headcount, at a cost per interaction fundamentally different from the fully loaded cost of a human advisor. Section 8.9 details this economics.

8.5 The channel becomes part of the sale: web, WhatsApp, and voice

Web, WhatsApp, and phone are three different entry points to the same need. On the web, a person arrives from a search, a campaign, or a specific page and expects an answer while intent is still active. A Meta ad can also open a WhatsApp conversation directly: Meta describes "click-to-WhatsApp" ads as Facebook or Instagram ads that send the person straight into a conversation with the business.

WhatsApp has become a major commercial channel. Meta says more than 3 billion people use WhatsApp across more than 180 countries, and more than 2 billion use it every day. In June 2026, Meta also reported more than one billion active conversations with businesses every day across WhatsApp, Messenger, and Instagram, and more than one million businesses already using Meta Business Agent on WhatsApp and

Messenger. For the enterprise, the benefit is straightforward: a click can become a conversation immediately, without forcing an intermediate form.

Voice completes the system. The WhatsApp Business Calling API can, among other things, receive user-initiated calls. In this model, the main focus remains inbound: a person clicks, writes, or calls because the need exists now. The AI agent must understand that request, mobilize enterprise knowledge, and respond before handing off to a human if necessary. Email and SMS then play complementary roles for confirmations, documents, links, appointments, or follow-up.

Official sources: Meta, "Click to Message Ads" and "Create Ads that Click to WhatsApp"; Meta, "Ways To Manage Your Businesses Chats On WhatsApp," April 2025; Meta, "Be There for Every Customer With Meta Business Agent," June 2026; Meta for Developers, WhatsApp Business Calling API.

8.6 The third shift becomes operational: consent, provenance, and inbound demand

The third shift introduced in the opening now takes an operational form. The source of the request, consent, the channel used, and the ability to preserve evidence of that consent become elements of the sales architecture, not merely separate legal issues.

In France, Law No. 2025-594 of June 30, 2025 shifts telephone solicitation, effective August 11, 2026, from an opt-out system - the Bloctel list - to a prior-consent system. Consent must be freely given, specific, informed, unambiguous, and revocable; it is valid for no more than one year, and proof must be retained by the professional for at least three years. Implementing Decree No. 2026-662 of July 23, 2026 specifies how consent must be collected, retained, and withdrawn. Violations of Articles L. 223-1 through L. 223-5 may result in administrative fines of up to EUR75,000 for an individual and EUR375,000 for a legal entity. Article L. 223-1 also provides that any contract concluded following telephone solicitation conducted in violation of its provisions is void. Separately, where abuse of vulnerability or ignorance is established under Article L. 132-14-1, penalties rise to five years' imprisonment and a EUR500,000 fine; the fine may be increased to 10% of average annual revenue. Bloctel ceases to exist on the same date.

At the European Union level, the situation requires a precise reading rather than a direct extrapolation from France. The ePrivacy Directive (2002/58/EC) already requires prior consent in all 27 member states for marketing by email, SMS, and automated calling systems. For telephone calls made by a human, however, the same directive allows each member state to choose between an opt-out system and a prior-consent system. The proposed ePrivacy Regulation, which could have harmonized this point across the EU, was formally withdrawn by the European Commission on July 16, 2025 because the co-legislators did not reach agreement. France's shift is therefore a national choice, within a broader movement already underway to varying degrees in several European countries, but no single binding EU regulation is currently on the Brussels agenda.

In the United States, the recent trajectory moved in the opposite direction. In late 2023, the Federal Communications Commission adopted a so-called "one-to-one consent" rule: consent had to be given to a specifically named seller rather than shared among multiple marketing partners. Its intent was similar to the French reform.

That rule did not survive. The U.S. Court of Appeals for the Eleventh Circuit vacated the restrictions on January 24, 2025, finding that the FCC had exceeded its authority. The FCC then removed those restrictions from its rules through a final rule published in the Federal Register on August 29, 2025.

The applicable U.S. framework therefore remains the prior regime: prior express written consent is still required for advertising or telemarketing robocalls and robotexts covered by the TCPA, particularly when

they use an automatic telephone dialing system or an artificial or prerecorded voice. The National Do Not Call Registry continues to apply in parallel to telemarketing solicitations within its scope. Structurally, this framework is closer to France's former Bloctel regime than to the new generalized opt-in regime.

The strategic consequence remains similar on both sides of the Atlantic, although for different reasons. In France and several neighboring European markets, regulatory constraints mechanically reduce the volume of outbound prospecting that can be conducted, shifting value toward the ability to capture and convert inbound demand. In the United States, where the restriction did not tighten in the same way, the argument is primarily competitive: response speed already determines conversion, as shown by the data in Sections 8.4 and 8.8, regardless of any future regulatory change.

Sources: Légifrance, Law No. 2025-594 of June 30, 2025 and Decree No. 2026-662 of July 23, 2026; economie.gouv.fr, DGCCRF; European Commission, withdrawal of the proposed ePrivacy Regulation, July 16, 2025; Insurance Marketing Coalition v. FCC, Eleventh Circuit, January 24, 2025; FCC, final rule, September 2025.

FROM ANALYSIS TO APPLICATION

The preceding sections describe the market. The following sections explain how SACHA OMON AI responds to it.

8.7 Anatomy of a multichannel agentic AI SDR agent

A multichannel agentic AI SDR agent applies the architecture described in Chapter 7 to the top of the sales cycle. Agentic reasoning drives the conversation and action; RAG ensures that each statement made to the prospect comes from the enterprise's validated knowledge.

DIAGRAM 12

The customer journey of a multichannel agentic AI SDR agent

Seven stages; at each stage, what agentic reasoning drives and what RAG contributes.

STEP	WHAT THE AI AGENT DOES	WHAT RAG PROVIDES
1 • Capture	Detects interest, the initial need, and the channel.	Open the conversation on the right channel and in the right language.
2 • Intent detection	Detects the project, urgency level, and intent.	Validate the product, need, and use case.
3 • Qualification	Qualifies the request, criteria, and profile.	Eligibility criteria, cases, blockers, and geographic areas.
4 • Objection handling	Addresses objections, reassures, and reframes.	Validated arguments and knowledge-grounded answers.
5 • Appointment booking	Proposes a time slot, confirms it, and collects information.	Assignment rules, calendar, and priorities.
6 • Handoff	Creates the record and hands it to sales.	Structured summary, context transmitted to the CRM.
7 • Follow-up	Follows up when needed and maintains engagement.	Follow-up rules, scenarios, and controlled cadence.

Figure 12: The customer journey of a multichannel agentic AI SDR agent, step by step.

CYCLE STAGE	CONTRIBUTION OF AGENTIC REASONING	CONTRIBUTION OF RAG
Capture	Immediate engagement on the visitor's chosen channel, at any time.	Opening response contextualized by the offer and the page viewed.
Intent detection	Reads buying signals and adapts the question path.	Understands product vocabulary and documented use cases.
Qualification	Adaptive question sequence rather than a fixed form.	Eligibility criteria, pricing constraints, and geographic rules retrieved from internal sources.
Objection handling	Reframing, targeted follow-up, handling uncertainty.	Validated talking points, references, guarantees, and terms cited from the source.
Appointment booking	Calendar lookup, proposal, confirmation, reminder.	Assignment rules by territory, expertise, and availability.
Handoff	Create the CRM record, human handoff at a defined threshold.	Structured and sourced conversation summary.
Follow-up	Calibrated follow-up cadence on the most engaging channel.	Complete contact history across all channels.

Table 33: Division of roles between agentic reasoning and grounded knowledge across the top of the sales cycle.

8.8 The metrics to instrument

METRIC	DEFINITION	MARKET REFERENCE POINT
First-response time	Time between receipt of the request and the first useful response.	Project target: under 60 seconds. Historical reference: in the 2011 Harvard Business Review audit (hbr.org/2011/03/the-short-life-of-online-sales-leads) covering 2,241 U.S. companies, the average response time was 42 hours among companies that responded within 30 days.
After-hours capture rate	Share of requests handled outside human operating hours.	Structural opportunity: approximately 75% of calendar time.
Autonomous qualification rate	Share of conversations qualified without human intervention.	Measure on the actual project scope; this White Paper does not use a single cross-market benchmark.
Kept-meeting rate	Meetings attended divided by meetings booked.	A qualification-quality metric that is more reliable than raw volume.
Source faithfulness	Share of answers fully supported by the knowledge base.	Measured continuously through model evaluation and human review.
Escalation rate	Share of conversations handed off to a human.	A stable, intentional rate is better than a zero handoff rate.
Cost per qualified meeting	Fully loaded cost divided by qualified meetings generated.	Track by channel, conversation type, and autonomy level. The objective is to measure system performance, not to manufacture an artificial comparison with a human job.

Table 34: Minimum dashboard for a multichannel agentic AI SDR agent.

8.9 The economics of an agent: measure real cost, without an artificial ceiling

There is no universal price or meaningful cost ceiling for an AI agent. Cost depends on the model selected, token volumes, voice, channels, RAG, tools called, integrations, infrastructure, supervision level, and the degree of autonomy entrusted to the system.

The right economic approach is to track the system's actual costs and connect them to useful outcomes: conversation handled, request qualified, meeting held, issue resolved, action completed, or attributable revenue. This measurement must evolve with the architecture because models, platforms, and development methods are changing rapidly.

PRINCIPLE

Measure what the agent costs and what it produces, without pretending that a price range or a human-versus-machine comparison summarizes the economics of a project.

APPLICATION: SACHA OMON AI

SACHA OMON AI is a multichannel agentic AI SDR agent: one AI agent that acts from the moment an inbound request arrives across website, WhatsApp, phone, email, and SMS, with centralized context across all channels, 24/7 and in more than 40 languages.

Its architecture applies the pattern described in this White Paper: the Sacha Engine agentic core drives reasoning, purchase-intent detection, and qualification logic; a dedicated RAG knowledge base grounds every response in customer-validated documentation; a single-tenant infrastructure isolates each customer's data, with European or U.S. hosting; connected operations link the AI agent to the CRM, calendar, human handoff, and dashboard.

The AI SDR agent is operated, maintained, and updated by SACHA OMON AI, allowing the customer to focus its teams on converting meetings rather than operating a technical platform.

This interoperable architecture - dedicated RAG and open rather than proprietary protocols - also positions the company for the market's next stage, when a customer's personal agent will communicate directly with the company's agent (Section 2.5).

8.10 From tokens to the unit cost of an agentic task

Tokens are one of the core economic units of LLMs: providers generally charge for input and output volume, with pricing that varies by model, caching, context, and reasoning level. In an AI agent, a single request may trigger multiple model calls, so token management becomes a design and operating issue.

Even so, an agent's cost is not limited to tokens. Depending on the case, it also includes retrieval and RAG, tool calls, compute, storage, voice or telephony, channels, and supervision. The metric that matters to leadership is the fully loaded cost per useful outcome: qualified conversation, resolved request, prepared meeting, or completed action, compared with the business value produced.

This economics explains why calibration matters. A well-designed agent uses only the reasoning level required, avoids unnecessary loops, and reserves the most expensive models for situations that justify them. Digital production cost can then be compared with human time saved, additional availability, service quality, and revenue protected or created.

8.11 One intelligence layer, five entry points

A sales or service agent should not become a collection of separate bots. A prospect who starts on the web, continues on WhatsApp, receives an email or SMS, and then calls is experiencing one relationship. The architecture must therefore unify identity, history, rules, and the next action while adapting expression to each channel.

The journey follows six consistent steps. The channel connector turns the message, voice, or attachment into a usable event. The system links that event to the correct contact and retrieves authorized context. The agentic engine identifies intent and decides whether to answer, search, ask for clarification, or act. RAG supplies the validated passages. The business tool performs any required action. Finally, the response is reformulated for the channel and the trace is retained for the next interaction or human handoff.

STEP	QUESTION HANDLED	EXAMPLE
1. Normalize	What just happened on this channel?	Transcribe a call or read a WhatsApp attachment
2. Identify	Who is speaking and what context may be used?	Match the phone number to the correct CRM contact
3. Reason	What is the intent and the next useful step?	Distinguish a question, an objection, and an urgent request
4. Ground	Which authorized sources answer the request?	Retrieve the price, rule, and validity date
5. Act	What effect should be produced in the systems?	Create an opportunity or book a time slot
6. Hand off	How can the conversation continue without losing context?	Send the summary, sources, and next action to the advisor

Table 35: The same agentic RAG cycle behind every channel.

On WhatsApp, agentic AI is especially useful because the conversation is discontinuous. A prospect may reply hours later, send a photo, a voice note, or a simple confirmation. External memory preserves the state of the journey; RAG retrieves the exact information; the agent decides whether it can continue or should hand off. Messaging becomes a durable service environment rather than a session-based chatbot.

On the web, speed and browsing context dominate. The page being viewed, source campaign, and observed behavior can help determine the first questions, subject to the consent framework. RAG prevents the agent from improvising an offer; calendar or CRM integration can immediately turn intent into a sales step.

On the phone, the main constraint is latency. The voice, transcription, reasoning, retrieval, and speech-synthesis chain must respond in real time, handle interruptions cleanly, and avoid overly long phrasing. The knowledge remains the same, but the expression engine selects shorter answers. As soon as strong emotion, complex negotiation, or risk appears, the call can be transferred with the transcript, reason for escalation, and already verified information.

8.12 Pre-sales and after-sales: two complementary entry points

Pre-sales is a natural entry point because it combines high value with repeatable tasks: understanding the need, answering questions, comparing options, handling simple objections, checking eligibility, gathering useful information, and preparing a meeting. The agent does not replace the sales representative; it hands over a request that is understood, documented, and already qualified, at the moment human involvement creates the most value.

After-sales is equally important. A RAG knowledge base can bring together manuals, procedures, equipment references, maintenance instructions, error codes, FAQs, and warranty rules. The customer can receive a useful first response without waiting for service hours or repeating the problem to multiple people.

On WhatsApp, the use case becomes very concrete: a customer can send a photo of equipment, a control screen, an error code, a part, or a diagram. A model capable of image analysis can describe what it sees; RAG can then retrieve the corresponding validated manual or procedure; the agent can provide the first authorized checks while preserving the context of the exchange.

For heat-pump maintenance, for example, the agent can help identify a displayed code, recall a documented maintenance step, verify a manufacturer-approved setting, or prepare a first-level diagnosis. If the situation is ambiguous, an action presents risk, or physical intervention is required, the system alerts and sends the technician the photo, history, checks already performed, and sources consulted.

The benefit is therefore not to eliminate after-sales service. It is to filter simple requests, provide an immediate answer when it is safe and documented, and reserve technicians' time for cases that truly require their expertise. The same knowledge can support pre-sales, installation, and service years after the sale.

MOMENT	WHAT THE AGENT PREPARES	WHAT THE HUMAN OWNS
Discovery	Initial questions, context, relevant documentation	Strategic listening and complex reframing
Qualification	Criteria, consistency, availability, summary	Exception arbitration and relationship potential
Proposal	Data, options, calculations, and supporting documents	Negotiation, commitment, and commercial creativity
Customer onboarding	Reminders, procedures, structured collection	Change management and relationship
Support	First-level diagnosis, sources, ticket	Expertise, empathy, and resolution of non-standard cases
Retention	Signals, history, suggested actions	Political understanding and account development

Table 36: A division of roles designed to expand service capacity.

8.13 The AI agent provides availability. Humans retain judgment and the relationship.

The most visible failures of AI in customer relationships have often started with a promise of substitution. That promise creates scopes that are too broad, handoffs that happen too late, and metrics focused on headcount reduction. A complementary strategy starts with the opposite question: what capabilities does the organization want to offer that it cannot currently provide consistently?

Those capabilities may include 24/7 availability, an initial response in every important language, perfect consistency on commercial terms, systematic case preparation, consent-based follow-up, or handoff without loss of context. Humans remain responsible for judgment, exceptions, empathy, negotiation, and the trust relationship. The agent takes on availability, memory, retrieval, reliable repetition, and bounded execution.

This framing also makes team adoption easier. The system is no longer presented as an observer designed to measure or replace employees, but as service infrastructure that eliminates blind spots and delivers better-prepared situations. Employees then become the main contributors to the quality of the knowledge base, escalation rules, and evaluation sets.

LEADING METRIC

First measure the added service capacity: requests covered outside business hours, response consistency, quality of transmitted context, first-contact resolution, satisfaction, and conversion. Time saved remains useful, but it should not be the only value story.

8.14 Application: SACHA OMON AI architecture

DIAGRAM 13

SACHA OMON AI functional architecture

Every layer interacts with the others; governance applies across the entire architecture.

1 · CHANNELS - CONVERSATION INPUT AND OUTPUT	Website · WhatsApp · Phone · Email · SMS. Two-way interactions: the agent engages users across all channels.
2 · ORCHESTRATION - REASONING, PLANNING, TOOL SELECTION, REFLECTION	Agent: agentic reasoning engine. Understands intent · plans · chooses the action · guides the journey.
3 · KNOWLEDGE AND MEMORY - WHAT THE AGENT KNOWS AND RETAINS	Knowledge (RAG): vector database · lexical index · graph · business connectors. Memory: session · contact profile · multichannel history. The agent queries to understand, then enriches and updates.
4 · TOOLS - ACTUAL EXECUTION OF ACTIONS	CRM · Calendar · Billing · Telephony · Payments. The agent selects and activates tools; it receives results, statuses, and returned data.
5 · GOVERNANCE - SECURITY, COMPLIANCE, AND QUALITY ACROSS ALL FOUR LAYERS	Logging · guardrails · continuous evaluation · human escalation. Humans retain accountability, judgment, and decision authority.

Cross-cutting governance: it continuously supervises, protects, and improves the entire architecture. Payment illustrates a future capability, not a feature currently offered.

Figure 13: Public functional architecture of SACHA OMON AI. Sacha Engine's internal mechanisms remain confidential.

SACHA OMON AI applies this model as a multichannel agentic AI SDR agent, deployed and operated for the customer. Sacha Engine provides proprietary orchestration and can use multiple LLM families according to the company's policy. Models provide understanding, reasoning, and formulation; knowledge, accounts, and operational assets remain governed within the customer's perimeter.

The White Paper intentionally does not disclose Sacha Engine's routing logic, context construction, memory, CRAG, control, or execution mechanisms. The public principle is simple: SACHA OMON AI understands the request, mobilizes authorized knowledge, prepares the useful action, and transfers the full context to a human whenever human involvement is required.

The service operates across web, WhatsApp, phone, email, and SMS, 24/7, in more than 40 languages. On real-time channels, the design target is a first useful response in under two seconds when the technical conditions of the channel allow it. The customer does not have to operate a technical platform: SACHA OMON AI configures, deploys, operates, maintains, and improves the system.

REFERENCE ARCHITECTURE

Single-tenant infrastructure: data hosted for the customer, managed operations, and a dedicated orchestration layer. The differentiator is not multichannel alone, but its combination with dedicated RAG, centralized context, deep business customization, and operations performed for the customer.

8.15 What the enterprise should retain as assets

A well-designed agentic project should not amount to a service that simply disappears when the contract ends. It should leave the enterprise with reusable and portable assets. That requirement protects the investment, makes it easier to change providers or models, and allows AI to expand into new functions without starting from zero.

ASSET	WHY IT MATTERS	EXPECTED CONTROL
Knowledge base	Captures sources, rules, and approvals	Exportable, documented, with identified owners
Accounts and channels	Preserves numbers, identities, and audiences	Opened and owned in the customer's name
Relationship data	Preserves history and customer continuity	Stored according to the customer's policies and rights
Rules and guardrails	Formalizes business operations	Readable, versioned, and auditable
Evaluation sets	Makes every change measurable	Reusable regardless of model
Execution traces	Explain sources, tools, actions, and escalations	Accessible with a defined retention period

Table 37: Assets that should remain usable by the enterprise beyond any one provider or model.

PART III

HOW TO DEPLOY

A. Frame your project - seven structuring questions, a standalone script, a method, and fourteen decisions.

This White Paper argues that an enterprise should remain in control of its knowledge, data, and accounts. It would be inconsistent to sell you a method that you could apply only by depending on us. This chapter therefore shows how to frame your project and draft a first version of your project requirements document with your own AI before taking any other step.

Seven structuring questions. Your starting point.

An AI-agent project does not begin with choosing a model or comparing platforms. It begins with seven structuring answers. They determine scope, the knowledge to mobilize, channels, rules, sizing, and the build approach. The standalone script on pages 89 to 92 adds an opening question - "In one sentence, in your own words, what do you want this agent to do?" - before returning to the seven questions. The script therefore asks eight questions in total.

1	Which markets will this AI agent serve? Determines languages, dominant channels, local rules, and regional architecture.
2	What industry do you operate in? Determines regulatory obligations, business vocabulary, and authoritative sources.
3	What is the first intended use case? Determines scope, the expected measurable outcome, and what the agent must hand off to a human.
4	Where does the knowledge the AI agent will need live today? Determines knowledge-base work, access rights, and RAG configuration.
5	How many requests of this type do you handle each month? Determines sizing, the true cost of a useful outcome, and the expected return.
6	How large is your organization? Determines governance, approval levels, and deployment pace.
7	Who will build it? Determines the division of roles: internal team, systems integrator, or end-to-end managed service.

Answer these seven questions and you have the foundation of your project.

WHAT YOUR OWN AI DOES WITH YOUR ANSWERS

From your answers to a project requirements document with your own AI

These seven questions structure the project. To turn the answers into a usable document, the standalone script on pages 89 to 92 adds an opening question - "In one sentence, in your own words, what do you want this agent to do?" - then returns to the seven previous questions. It therefore asks eight questions in total, one at a time, in your own AI - ChatGPT, Claude, Gemini, or Grok - before drafting a first version of the project requirements document.

No external access is required: you copy the script contained in this White Paper and paste it into the AI you already use. The resulting document remains a working hypothesis to be corrected and validated by your teams.

YOU GET	WHAT IT IS	WHAT YOU DO WITH IT
The 14 decisions, pre-drafted	A first draft of the 14 requirements-document decisions, adapted to your answers	You revise it with your teams instead of starting from zero
The method applied to your situation	The 8 deployment rungs, sequenced around your first use case	You know where to start and in what order
A target architecture	The required building blocks and the technical points to validate	You frame the discussion with your IT team or service provider
Cost assumptions	Orders of magnitude for the cost of the useful outcome, not just the model	You prepare a defensible budget
Rules to verify	The regulatory points applicable to your markets and the questions to ask providers	You de-risk the project before committing

Table 38: What the standalone script produces from your eight answers.

None of these five outputs is a contractual commitment. They are working hypotheses to be discussed, corrected, and validated by the enterprise.

WHAT HAPPENS NEXT · SACHA OMON AI METHOD

Stair-step deployment: do not build everything; build the first rung

Once the framing is in place, the next question is pace. The method advocated in this White Paper is pragmatic: design the overall architecture from the start, but do not spend months freezing every building block of a system. Models, platforms, and tools evolve too quickly. Each rung produces evidence and prepares the next one.

01	Real need Choose a concrete problem, an expected outcome, and a bounded scope.
02	Real-world proof Put a first capability in front of real cases and measure what works.
03	Expansion Correct the right layer, incorporate technological advances, and add the next rung.

The eight operational rungs

1. Choose a real, measurable need.
2. Define the authoritative knowledge and rules.
3. Choose a first channel and a bounded scope.

4. Give the AI agent a first useful capability.
5. Expose it quickly to real cases, under control.
6. Measure quality, delay, actions, escalations, and real cost.
7. Fix the right layer: data, rules, tool, model, or process.
8. Add the next rung while incorporating technological advances that appeared in the meantime.

FOUNDER'S CONVICTION

The objective is not to build the AI agent of 2028 today. It is to build the first rung today of an architecture capable of getting there.

This trajectory is the method proposed by Salomon Bellaïche in this White Paper. It is not a universal standard: the right pace depends on each organization's risk, business, data, and responsibilities.

CONTROL GRID · THE 14 DECISIONS

The 14 decisions to document - 1 through 7

Here is the complete grid. The standalone script allows your own AI to pre-draft a first version from your eight answers. The grid then serves as a control: until a decision has evidence beside it, it has not been made.

DOMAIN	PRACTICAL QUESTION	EXPECTED EVIDENCE
Objective	What concrete outcome should the AI agent achieve?	Measurable, accountable outcome identified
Scope	Which requests does it handle, and which does it hand off?	Included and excluded cases, escalation rules
Knowledge	Which sources are authoritative?	Inventory, owners, update dates
Permissions	What may it see, and for whom?	Access controls all the way through retrieval
Channels	Where does the request actually begin?	Web, WhatsApp, phone, email, SMS: volumes and constraints
Memory	What context should persist from one interaction to the next?	Retention policy and contact identity
Tools	Which actions may the AI agent actually execute?	Allowlist, approval, reversibility, and logging

Table 39: The first seven decisions in an AI-agent project requirements document.

YOU CAN WRITE THEM YOURSELF - OR HAVE AI PRE-DRAFT THEM

The standalone script turns your eight answers into an initial hypothesis for each of these 14 decisions. Your job is then not to start from zero, but to revise it with your teams.

HOW TO DEPLOY · THE 14 DECISIONS

The 14 decisions to document - 8 through 14

DOMAIN	PRACTICAL QUESTION	EXPECTED EVIDENCE
Human	When should the AI agent hand off?	Thresholds, routing, and transmitted context
Quality	How do we know it is answering and acting correctly?	Real cases, thresholds, recurring evaluations
Security	How do we limit injection attacks, leaks, and unauthorized actions?	Isolation, filters, testing, and traces
Portability	Who owns the accounts, numbers, data, and rules?	Ownership, export, and reversibility
Economics	What is the true cost of the useful outcome?	Models, voice, tools, infrastructure, operations, and oversight
Deployment	Who configures, operates, corrects, and evolves the system?	Responsibilities, service levels, and change process
International	Which languages, regions, and regulatory regimes are involved?	Regional architecture, localization, and market-specific rules

Table 40: The final seven decisions in an AI-agent project requirements document.

Pages 87 and 88 form the control grid for the project requirements document. The standalone script allows your own AI to prepare a first hypothesis from your eight answers; the enterprise must then verify sources, rights, actions, security, economics, and human responsibilities.

FROM THE GRID TO THE REQUIREMENTS DOCUMENT

You now have an 8-rung method and a 14-decision framework. The standalone script below turns this framing into a first version of a requirements document using your own AI. Chapter 9 then begins on page 93: "How to build and deploy an AI agent in the enterprise." Understand -> Start -> Build.

HOW TO DEPLOY · STANDALONE PROJECT REQUIREMENTS DOCUMENT

Have your own AI draft your project requirements document

The principle is simple. Copy the boxed script below and paste it into your own AI - ChatGPT, Claude, Gemini, or Grok, whichever you already use. It will ask you eight questions, one at a time, and wait for your answer before asking the next.

At the end of the interview, it drafts your project requirements document: the 14 documented decisions, the final question, the 8 deployment rungs applied to your situation, an initial estimate of the real cost per useful outcome, and the 5 questions to ask any vendor from Chapter 9.4. Ten to fifteen minutes is enough, with no preparation and no additional subscription: the script uses the AI you already have.

This White Paper argues that an enterprise should remain in control of its knowledge, data, and accounts. It would be inconsistent to sell you a method that you could apply only by depending on us.

The script begins below and continues on the following pages. Copy it in full without changing a word, then paste it into a new conversation with your AI.

THE SCRIPT · COPY WORD FOR WORD

You are an experienced project manager specializing in RAG-augmented agentic AI agents. You apply the SACHA OMON AI method published in the White Paper "How AI, powered by agentic AI and RAG, is now entering the enterprise."

YOUR ROLE

Conduct a short interview, then draft a project requirements document. You are neither a salesperson nor an evangelist. You are rigorous and practical, and you say when something cannot yet be decided.

REQUIRED PROCESS

Ask the eight questions below ONE AT A TIME. After each question, stop and wait for my answer. Never ask two questions at once. Do not draft anything before the eighth answer. If an answer is vague, ask one follow-up question, then move on.

THE EIGHT QUESTIONS

1. In one sentence, in your own words, what do you want this agent to do?
2. In which countries or regions are your users located?
3. What industry are you in?
4. What is the first intended use case? One only.
5. Where does the knowledge the agent will need live today? (SaaS tools, scattered files, structured database, paper or PDF, in employees' heads, a mix of everything)
6. How many requests of this type do you handle per month?
7. How large is your organization?
8. Who will build it? (internal team, agency or systems integrator, turnkey vendor, a mix, not decided yet)

AFTER THE EIGHTH ANSWER, PRODUCE THE FOLLOWING DOCUMENT

A. THE 14 DECISIONS TO DOCUMENT

For each one, write four blocks in this order:

- IN PLAIN ENGLISH - one simple sentence understandable to a nontechnical executive.
- OUR PROPOSAL - one paragraph written for MY situation, using my words.
- EVIDENCE REQUIRED - what I should be able to show.
- WHY IT MATTERS - one sentence, without hype.

1. Objective - What concrete outcome must the agent achieve?

Evidence: measurable outcome and identified owner.

2. Scope - Which requests does it handle, and which does it hand off?

Evidence: included/excluded cases and escalation rules.

3. Knowledge - Which sources are authoritative?

Evidence: inventory, owners, update dates.

4. Access rights - What can it see, and for whom?

Evidence: access controls enforced all the way through retrieval.

5. Channels - Where does the request actually begin?

Evidence: web, WhatsApp, phone, email, SMS - volumes and constraints.

6. Memory - What context must persist from one interaction to the next?

Evidence: retention policy and contact identity.

7. Tools - What actions can the agent actually execute?

Evidence: allowlist, approval, reversibility, logging.

8. Human handoff - When must the agent hand over to a person?

Evidence: thresholds, routing, transmitted context.

9. Quality - How do we know it is answering and acting correctly?

Evidence: real cases, thresholds, regular evaluations.

10. Security - How do we limit prompt injection, data leakage, and unauthorized actions?

Evidence: isolation, filters, tests, traces.

11. Portability - Who owns the accounts, numbers, data, and rules?

Evidence: ownership, export, reversibility.

12. Economics - What is the true cost per useful outcome?

Evidence: models, voice, tools, infrastructure, operations, supervision.

13. Deployment - Who configures, operates, fixes, and evolves the system?

Evidence: responsibilities, service levels, change process.

14. International - Which languages, regions, and regulations are involved?

Evidence: regional architecture, localization, market-specific rules.

B. THE BASELINE

Ask me to record three numbers before we begin, and explain in one sentence why: without them, no one will be able to say in six months whether the project succeeded. Current average response time; monthly volume handled; monthly cost of this activity; date of measurement.

C. THE FINAL QUESTION

"If the agent stops working tomorrow, what remains with the enterprise?"

Answer for my situation. A good answer includes: an owned knowledge base, exportable data, documented rules, evaluation sets, accounts under company control, and a stronger human process. Tell me which ones I do not have.

D. THE EIGHT OPERATIONAL RUNGS

Architectural vision, stair-step deployment. Place each rung in MY situation:

1. Choose a real, measurable need.
2. Define the authoritative knowledge and rules.
3. Choose a first channel and a bounded scope.
4. Give the agent a first useful capability.
5. Expose it quickly to real cases, under control.
6. Measure quality, delay, actions, escalations, and real cost.
7. Fix the right layer: data, rules, tool, model, or process.
8. Add the next rung while incorporating advances that appeared in the meantime.

E. THE TRUE COST PER USEFUL OUTCOME

Based on my answers, estimate only what can reasonably be estimated. Give order-of-magnitude estimates for model cost per useful outcome and human time that could potentially be redeployed. For the share of requests the agent might resolve on its own once mature, do not give any percentage if my answers do not provide enough information about the nature and complexity of requests, exceptions, the condition of the knowledge base, and the actions required. In that case, state precisely what data is missing. If you offer a range, label it explicitly as a hypothesis and explain what it is based on. Show all assumptions. Remind me of the three cost items vendors often omit: initial knowledge-base preparation, monthly operating workload, and the cost of requests that will still be handled by teams.

F. THE FIVE QUESTIONS TO ASK ANY VENDOR

For each one: the exact question, the answer that should raise a red flag, and the answer you want to hear.

1. What exact sources does your agent rely on, and how can we verify that?
2. What actions does it actually execute, which require human approval, and where is the audit log?
3. How do you measure quality in production, and who owns the evaluation set?
4. In whose name are the accounts, keys, and numbers opened? What can I export, in what format, and how quickly?
5. How much internal time do you need, from whom, each month?

G. WHAT YOU CANNOT DECIDE

Finish by listing the decisions that cannot be finalized from eight answers, and explain why: they depend on my actual contracts, the real condition of my documents, my directory and identity systems, and my actual numbers. Be precise and candid. Do not fill in gaps.

ABSOLUTE RULES

- Never invent a legal rule, effective date, statutory provision, or regulation. For Decision 14, identify only the categories of obligations to verify for my countries (data protection, transparency around AI use, sector-specific rules, export controls if relevant), and state that dates and details must be verified with legal counsel or a current official source.
- Never call this "automation." Automation executes a rule written in advance; an agentic system makes decisions within a defined framework.
- Write in the language used by the person you are addressing (English, French, Hebrew, German, Italian, or another language as appropriate), in a restrained style, without hype, superlatives, or emoji.
- Never present a number as a commitment: these are order-of-magnitude estimates, and you must say so.
- If I ask you to skip steps, refuse once and explain what will be lost.

START NOW with Question 1 only.

BEFORE YOU START

Three precautions

1. The resulting document is a working draft, not a final document: it must be corrected with your teams.
2. Any numbers produced are order-of-magnitude estimates, never commitments.
3. On regulation, the script prohibits the AI from asserting a date or legal text: it identifies only the categories of obligations to verify with legal counsel.

The next chapter shows how to build and deploy this system inside the enterprise.

B. DEPLOY ACROSS THE ENTERPRISE

Build, govern, avoid mistakes - then conclude.

09

How to build and deploy an AI agent in the enterprise

*Think about the architecture
globally. Deploy it rung by rung.*

Successful integration does not begin by mapping every process, every dataset, and every system in the enterprise. That effort delays results, diffuses responsibility, and feeds pilots that never get adopted. Instead, choose a precise entry point where the company is already absorbing the most lost value - often marketing acquisition and inbound-demand handling. Produce visible proof on real traffic, then expand the scope one source, one channel, one tool, and one team at a time until a shared nervous system emerges.

9.1 The seven structuring decisions

DECISION	QUESTION TO RESOLVE	RECOMMENDATION BASED ON FIELD FEEDBACK
Scope	Which process, what volume, what measurable outcome?	A high-volume, low-variance process, with a financial metric identified during initial scoping.
Knowledge sources	Which documents, how fresh, with what permissions?	A narrow, perfectly maintained scope outperforms a broad, approximate one.
Model	Which reasoning engine, what latency, what cost?	Different models based on query difficulty; portability preserved by design.
Tools and integrations	Which systems does the agent actually write to?	Standardize on an open protocol; start with read access, then open write access in stages.
Autonomie	Where is the boundary between agent and human?	Reversible actions delegated to the agent, irreversible actions subject to approval.
Evaluation	How is quality measured continuously?	Build a business test set before development and run it at every change.
Ownership	Who owns the accounts, data, and phone numbers?	Accounts and data in the customer's name, with no dependency created by the solution.

Table 41: The seven decisions to make before the first line of code.

9.2 Start with a specific need, prove value, then expand

DIAGRAM 14

Start where value is already being lost

Four criteria for choosing the entry point, followed by six stages from analysis to centralization.

BEFORE YOU START - FOUR CRITERIA FOR CHOOSING THE ENTRY POINT

A	ACQUISITION A marketing spend that already exists.
B	DEMANDE ENTRANTE Prospects and customers who are already waiting.
C	ONE METRIC ONLY Response, qualification, or revenue - not all three.
D	NO ENTERPRISE-WIDE TRANSFORMATION A result the team can see.

THE PATH - SIX STAGES, FROM ANALYSIS TO CENTRALIZATION

01	TARGETED ANALYSIS · D1 -> D3 One journey, one loss point, one owner, one metric, and only the sources required.
02	FIRST SCOPE · D4 -> D7 Minimum validated RAG base, rules, escalation, and one pilot channel or campaign.
03	REAL TRAFFIC · D8 -> D14 Controlled go-live, logging, human takeover, and daily correction.
04	PROOF · D15 -> D30 Measure speed, quality, adoption, conversion, service, and cost.
05	PROGRESSIVE EXPANSION · AFTER PROOF One source, one channel, one tool, one team at a time.
06	CENTRALIZATION · ONGOING Knowledge, context, permissions, traces, and evaluations brought together in Sacha Engine.

Exit principle: each extension must preserve the measured gain, response quality, and the ability for humans to take over.

Figure 14: Recommended path - four criteria for choosing where to start, followed by six stages detailed in Table 42 on the next page.

PHASE	DURATION	DELIVERABLES	EXIT CRITERION
Targeted analysis	D1 -> D3	One entry point, one loss point, one owner, one metric, and only the strictly necessary sources.	Scope, owner, and success threshold validated.
First scope	D4 -> D7	Minimum validated RAG base, rules, escalation, and one pilot channel or campaign.	Business test cases passed under human oversight.
Real traffic	D8 -> D14	Controlled go-live, logging, human takeover, and daily correction.	Real requests handled end to end.
Evidence	D15 -> D30	Measure speed, quality, adoption, conversion, service, and cost.	Visible gain and an explicit decision to continue.
Progressive expansion	After proof	Add one source, one channel, one tool, or one team at a time.	Each expansion preserves the metric and quality level.
Centralization	Ongoing	Knowledge, context, permissions, traces, and evaluations brought together in Sacha Engine.	A shared, governed, reversible central nervous system.

Table 42: Rapid progression model. The Day 1-Day 30 windows assume that access, sources, and business approvals are available; they describe a first entry point, never an enterprise-wide deployment.

9.3 The clean-data rule

One finding runs through all the experience reports reviewed: data preparation consumes more time than deploying the agent itself. A 2026 industry survey ranks data quality as the number-one deployment barrier, cited by 52% of respondents. The most detailed case studies explicitly report preliminary cleanup phases that took longer than installation.

The practical consequence belongs at the top of the project plan: document-preparation budget is a performance investment, and underestimating it is the leading cause of the gap between promised and achieved results.

Source: compilations of 2026 enterprise surveys; vendor case studies.

9.4 The five questions to ask any vendor

1. What sources does the agent base its answers on, and how do you verify that? The answer you want describes citation mechanisms and faithfulness measurement rather than a general claim of quality.
2. What actions does the agent actually execute in our systems? The difference between "answering" and "writing to the CRM and booking a meeting" determines the value produced.
3. How is quality measured on our traffic after go-live? Continuous evaluation on real data is the best predictor of sustained performance.
4. Who owns the accounts, numbers, and data? Customer ownership protects reversibility and the value of the asset created.

5. What internal resources does the system require? A vendor-operated AI agent frees internal teams to focus on conversion rather than technical operations.

9.5 Start with a visible problem, then reuse the same knowledge foundation

An enterprise can enter agentic AI more easily through a visible problem than through a broad transformation program. In digital acquisition, the marketing investment already exists and the loss can be measured: unanswered forms, callback delay, abandoned questions, unqualified requests, or context lost between channels. This entry point provides immediate proof without forcing teams to redesign the entire company. Once adopted, the same foundation of knowledge, memory, rights, and measurement can support after-sales, customer support, operations, and other functions.

STEP	DURABLE DELIVERABLE	PROGRESSION CRITERION
1. Analyze	Entry point, cost, volumes, owner, and metric	Problem and success threshold accepted
2. Ground	Minimum sources, permissions, RAG base, and test cases	Relevance and faithfulness at the defined threshold
3. Serve	One channel and one journey with clear value	Stable quality on real requests
4. Prove	Business measurement, adoption, and human takeover	Visible gain and decision to continue
5. Expand	Additional sources, tools, languages, channels, and teams	Each expansion maintains the outcome
6. Centralize	Sacha Engine: context, governance, traces, and evaluations	Shared and reversible central nervous system

Table 43: A cumulative roadmap: each stage creates a reusable asset.

The first scope should be important enough to create value but bounded enough to evaluate. A pre-sales or after-sales knowledge base works well: questions are numerous, sources exist, human handoff is natural, and gains can be measured without immediately giving the agent irreversible decision authority.

9.6 Measure augmented service, not productivity alone

The dashboard for an agentic RAG system must connect several levels. Retrieval quality measures whether the right sources are found. Response quality checks faithfulness and usefulness. Execution quality controls tools and task completion. Service quality observes what customers and human teams experience. A single metric cannot diagnose a performance decline.

LEVEL	RECOMMENDED METRICS	MANAGEMENT QUESTION
Knowledge	Retrieval precision and recall, freshness, coverage	Does the agent find the right evidence?
Response	Faithfulness, citations, abstention, consistency	Does it say only what the sources support?
Action	Tool success, completion, error, reversibility	Was the intended outcome actually produced?
Experience	P50 and P95 latency, satisfaction, repetition, abandonment	Is the service smooth for the user?
Human	Handoff quality, takeover time, complete context	Does the team receive a better case file?
Business	Conversion, resolution, revenue, cost per conversation	Does the system create value?

Table 44: The six levels to connect in an agentic RAG system dashboard.

9.7 When an answer is wrong, fix the right layer instead of changing models

Measurement becomes useful when it identifies the layer responsible. An incorrect answer may come from a missing source, poor chunking, imprecise retrieval, an ambiguous response rule, or an unavailable business tool. Without this deeper diagnosis, the team may switch models when the real problem is data or process.

The monthly review should therefore combine real conversations, technical metrics, and feedback from the teams that take over cases. Selected corrections feed the knowledge base, escalation rules, and evaluation sets. The system improves through controlled organizational learning rather than relying on implicit model memory.

IMPROVEMENT LOOP

Observe the gaps, attribute the cause, correct the right layer, rerun the test cases, then measure again on real traffic. This discipline turns a one-off pilot into an enterprise capability.

10

Governance, security, and compliance

As soon as an agent acts in your systems, it becomes a full-fledged IT workload.

An agent has an identity, permissions, and access to data. Consent, request provenance, access rights, security, and evidence therefore stop being separate topics: they define the required level of control. This chapter describes the six controls to put in place, the frameworks applicable in the United States, the European Union, Israel, and Asia, and the risks specific to agents and how to mitigate them.

10.1 The six controls to put in place

CONTROL	PURPOSE	TYPICAL IMPLEMENTATION
Agent identity	Give the agent its own identity, separate from a user's account.	Enterprise directory, dedicated service accounts, credential rotation.
Action scope	Define deterministically what the agent may do.	Per-tool authorization policies, allowed-action lists, caps.
Mandatory grounding	Require a source for every factual claim.	Abstention instruction, systematic citations, documented refusal.
Logging	Preserve the trace of the plan, tool calls, sources, and decisions.	End-to-end observability, retention aligned with internal policy.
Continuous evaluation	Measure quality on real traffic at every change.	Business test sets, model evaluators, periodic human review.
Human escalation	Hand off at the right moment, with full context.	Explicit thresholds, conversational handoff, resume without repetition.

Table 45: Governance foundation for an agent in production.

2026 surveys put the share of organizations with a mature governance model for autonomous agents at about 21%. This gap between deployment and governance is the market's main risk factor and partly explains Gartner's forecast of project cancellations.

10.2 The United States: start with the market, govern on two levels

The U.S. market is not governed by a single equivalent of the EU AI Act. As of August 19, 2026, governance combines a federal environment favorable to innovation, existing sector-specific and consumer-protection

rules, the voluntary NIST framework, and state-level laws. For an enterprise serving the United States, the right question is therefore not only "Are we compliant at the federal level?" but "In which states, sectors, and decisions does our agent operate?"

The federal foundation. Executive Order 14179 of January 23, 2025, followed by Executive Order 14365 of December 11, 2025, place AI leadership and adoption at the center of national policy and seek a relatively light national framework. The NIST AI Risk Management Framework nevertheless remains the leading voluntary operational reference for mapping, measuring, governing, and managing risk; its profile dedicated to generative AI covers provenance, pre-deployment testing, security, privacy, and incident management.

The state level. The Texas Responsible Artificial Intelligence Governance Act took effect on January 1, 2026. Colorado adopted a revised framework in May 2026 for decision technologies affecting employment, housing, credit, insurance, healthcare, education, or essential services; its core obligations are scheduled to begin January 1, 2027. Other states add rules on transparency, data, biometrics, or consequential decisions. The architecture must therefore be able to activate jurisdiction-specific policies, logs, notices, and human thresholds without rebuilding the agent.

Official sources: White House, EO 14179 and national framework · NIST, AI RMF · Texas, HB 149 · Colorado, SB26-189.

10.3 The European Union: a horizontal risk-based framework

An organization established in the European Union, or whose system affects people located in the EU, operates under two complementary bodies of law. GDPR governs personal data; the AI Act governs AI-system uses, roles, and risks. Since August 2, 2026, the main general provisions and transparency obligations apply, while certain obligations for high-risk systems follow the 2027-2028 timetable published by the European Commission.

- GDPR. Lawful basis, purpose limitation, data minimization, notice to individuals, retention periods, security, processors, and international transfers. A conversational agent collects personal data by design: every channel, memory layer, and connector must therefore be included in the data-processing map.
- AI Act. Transparency when a person interacts with AI, documentation proportionate to risk level, data governance, traceability, and human oversight. For a sales or service agent, classification of the use case matters more than the model name: informing a prospect does not create the same obligations as a decision affecting employment, credit, insurance, or access to an essential service.
- Sovereignty and isolation. Regional hosting and tenant-level data isolation have become common requirements in European RFPs, particularly in regulated sectors.

A good practice emerges from observed deployments: clearly disclose the nature of the system in the first message, offer human handoff at any time, and log those elements. This transparency can also improve commercial performance; a Kantar study found that 79.3% of consumers look for proof of legitimacy before engaging in a conversation.

Official sources: GDPR, Regulation (EU) 2016/679 · European Commission, AI Act implementation timeline.

10.4 Israel, India, and Asia: distinct frameworks, adaptable architecture

Asian markets do not form a single regulatory bloc. Israel, India, Singapore, and Japan favor innovation-friendly frameworks supplemented by data-protection, cybersecurity, and sector-specific rules. China

applies more directly binding obligations to public-facing generative AI services. An international architecture must therefore separate the common foundation - identity, rights, evidence, logging, security, and human handoff - from policies activated by country, channel, and use case.

MARKET	FRAMEWORK AS OF AUGUST 19, 2026	IMPLICATION FOR THE COMPANY
Israel	Sector-based, innovation-friendly approach; privacy protection, cybersecurity, responsible-use guidance; signatory to the Council of Europe Framework Convention on AI.	Document purpose, protect data, secure infrastructure, and maintain human control without freezing innovation.
India	India AI Governance Guidelines published in November 2025; Digital Personal Data Protection Act and 2025 Rules for personal data.	Appropriate consent and notice, minimization, security, accountability, and localization assessed based on the service.
Singapore	Model AI Governance Framework; agentic AI-specific framework published by IMDA in January 2026.	Limit agent powers, test actions, maintain human accountability, and prepare for incidents.
Japan	AI Guidelines for Business, version 1.2 published by METI and MIC in April 2026; risk-management and shared-accountability approach.	Continuous evaluation, transparency, security, stakeholder rights, and lifecycle governance.
China	Binding administrative measures for publicly accessible generative AI services, deep-synthesis rules, and registration obligations for certain algorithms.	Verify public-facing scope, content, data, local providers, filings, and operations before launch.

Table 46: A common governance foundation, with local policies activated by market. This summary does not replace legal analysis of the specific use case.

Official sources: Israel, Responsible AI Use Guide · India, DPDP Rules 2025 · Singapore, Agentic AI Framework · Japan, AI Guidelines for Business · China, generative AI measures.

10.5 Risks specific to agents

RISK	DESCRIPTION	COUNTERMEASURE
Prompt injection	External content diverts the agent from its objective through a hidden instruction.	Treat all retrieved content as data, never as an instruction; isolate sensitive tools.
Data leakage	An agent exposes information beyond the requester's permissions.	Security filters at the index layer, with access control propagated through retrieval.
Knowledge-base drift	The knowledge base ages and the agent cites outdated information.	Scheduled synchronization, validity dates, alerts for obsolete documents.
Attribution error	In a multi-agent chain, determining the source of an error becomes difficult.	Per-tool attribution and detailed logging, as recommended in the evaluation literature.
Runaway costs	Reasoning loops multiply model calls.	Effort proportional to difficulty, per-conversation caps, unit-cost monitoring.

Table 47: Risks specific to agentic architectures and corresponding safeguards.

11

Mistakes to avoid before and during deployment

Eight observed anti-patterns, and the correction for each.

Klarna had to rehire people after turning too much over to AI. 11x was accused of inflating its numbers. In the field, sales AI still falls short of its promises more often than humans do. These failures do not mean agentic AI does not work; they reveal eight specific mistakes, corrected one by one in this chapter.

ANTI-PATTERN	SYMPTOM	CORRECTION
The pilot with no metric	The project demonstrates technical capability without a quantified commitment.	Set the financial metric and success threshold during initial scoping.
The neglected knowledge base	Approximate answers blamed on the model when the source is actually at fault.	Audit, clean, and assign editorial ownership to the knowledge base.
Full autonomy from day one	Fully agentic execution with no escalation, declining satisfaction.	Start supervised, expand only on quantified proof.
Volume as strategy	More sends, degraded sender reputation.	Prioritize qualification and document grounding.
Uniform sophistication	High latency and cost on simple questions.	Calibrate reasoning effort to actual difficulty.
Provider-created dependency	Accounts, phone numbers, and data held by the vendor.	Require account ownership and data portability.
The isolated agent	One agent per department, with no shared context.	Centralized context across channels and agents.
Measurement abandoned after launch	Quality silently drifts.	Continuous evaluation on real traffic, monthly review.

Table 48: Eight observed anti-patterns and how to correct them.

A Salesforce survey published in February 2026 of 1,050 IT leaders illustrates one of these issues: organizations operate an average of 12 agents, a figure expected to rise 67% in two years, and half of those agents operate in silos rather than coordinated systems. Orchestration therefore becomes a question before the number of agents does.

12

Where value is lost - and how it comes back

*Models, tools, and architectures do
not create value by themselves.*

It is their integration into enterprise processes, teams, and governance that turns potential into durable results. The three shifts described in the opening converge here: immediacy changes service expectations, personal agents move the interface between individuals and organizations, and consent, provenance, and evidence become operational data.

Generative AI has entered mainstream consumer use; structured enterprise integration is only beginning. The challenge is therefore not to distribute one more conversational tool, but to connect LLMs to knowledge, permissions, channels, tools, teams, and economic objectives, then build that capability step by step on real traffic.

1. Treat the model as a component. Performance comes from the architecture built around the model, as MIT research makes clear.
2. Ground every answer. An unsupported claim is a risk; a claim tied to a source is an asset.
3. Give the agent the ability to act. Value materializes when the agent writes into business systems and closes the task loop.
4. Let reasoning drive retrieval. Decomposition, routing, and iteration produce the accuracy gains measured in academic research.
5. Calibrate effort. Direct response for simple cases, iteration for complex ones: latency and cost are controlled through proportionality.
6. Invest in data before the agent. Document preparation determines outcomes more than vendor selection.
7. Make human escalation explicit. A stable, well-designed handoff rate protects the brand better than fully agentic execution.
8. Measure continuously. Source faithfulness, resolution rate, unit cost, and satisfaction should be monitored on real traffic, not only on a demo set.
9. Standardize connections. An open, collectively governed protocol reduces integration cost and lock-in risk.
10. Start where value is being lost. Inbound demand provides a directly measurable first field: prove value in a real scope, then extend the architecture.

FINAL WORD

The LLM understands and formulates. RAG provides the company's facts. Agentic AI guides the journey. Humans retain accountability, judgment, and the relationship. Integration begins at a precise entry point where value is already being lost; it demonstrates an outcome, expands progressively, then becomes a shared central nervous system. That is how AI stops being an isolated tool and becomes a full-fledged enterprise function.

Salomon Bellaïche - Founder of SACHA OMON AI, Editor-in-Chief of the White Paper.

Decision summary: what you need to retain and decide

This entire White Paper comes back to one question: when a request appears, does your enterprise act, or does it merely respond?

At the end of this White Paper, one conclusion stands out: generative AI has become mainstream, but structured integration inside the enterprise remains a minority practice. The decisive step is not distributing one more tool; it is connecting models to governed knowledge, rules, channels, tools, and teams to produce measurable outcomes.

Two capabilities bridge that gap. Agentic architecture enables the system to pursue a goal, prepare steps, call authorized tools, and hand off at the right moment. RAG grounds the response in verifiable, updatable enterprise knowledge. The LLM understands and formulates; RAG provides the facts; agentic AI drives the journey.

FOUNDER'S CONVICTION - THINK GLOBALLY, BUILD STEP BY STEP

I propose thinking about the architecture as a whole - knowledge, permissions, channels, security, tools, and humans - then deploying it rung by rung: real need, first capability, field test, measurement, correction, and expansion. This is a working method proposed in this White Paper, not a universal standard.

Fourteen numbers to frame the decision

METRIC	VALUE	SOURCE
Weekly ChatGPT users in 2026	900 M+	OpenAI, 2026
Organizations regularly using AI in at least one function	88 %	McKinsey, State of AI, November 2025
Organizations that have productionized at least one agentic system	≈ 23 %	McKinsey, State of AI, November 2025
Generative AI pilots with no measurable P&L impact	95 %	MIT NANDA, The GenAI Divide, 2025

METRIC	VALUE	SOURCE
Enterprise applications embedding a specialized agent by the end of 2026	40% (versus less than 5% in 2025)	Gartner
Projected global agentic spending in 2026	201,9 Md\$	Gartner
Standalone "agentic AI" market, 2026	approx. \$9-12 billion, CAGR above 40%	Fortune Business Insights, Precedence Research, MarketsandMarkets
RAG market 2026	≈ 3,3 Md\$, TCAC ≈ 42 %	Next Move Strategy Consulting
LLM-based enterprise systems using some form of RAG (order of magnitude)	≈ 60 %	Xu et al., "RAGOps: Operating and Managing Retrieval-Augmented Generation Pipelines," 2025 (arxiv.org/abs/2506.03401)
Agentic projects likely to be canceled by the end of 2027	> 40 %	Gartner
Growth in tasks executed on Amazon Bedrock AgentCore over 6 months	15x	AWS, June 2026
Agentforce ARR at Salesforce FY2026 close	\$800 million (+169% year over year)	Salesforce, FY2026 results
Israeli high-tech exports in 2025	approx. \$85 billion, or 58% of the country's exports	Israel Innovation Authority, 2026
Israel's position as a global technology-financing hub in 2025	4th worldwide, 1st outside the United States; 35% of investment directed to core AI	Israel Innovation Authority, 2026

Table 49: Fourteen reference metrics. Details and links in Appendix D.

Eight conclusions

1. The value gap comes from architecture, not the model. MIT research attributes pilot failures to shortcomings in contextual learning and process integration rather than model performance alone.
2. Agentic AI turns response into outcome. An agent plans, calls tools, writes into business systems, and closes the task loop: value shifts from text generation to work execution.
3. RAG is becoming a reference architecture for enterprise AI. It addresses the dual requirement of accuracy and traceability and turns document assets into operational assets.
4. Vendors are converging on a comparable architecture. Models remain replaceable; knowledge, rules, channels, rights, and orchestration become durable enterprise assets.
5. Inbound demand is a directly measurable use case. In both B2B and B2C, response speed, qualification, and continuity of context make it possible to handle intent at the moment it exists.
6. The knowledge base is the first durable AI asset. Structuring sources, permissions, and validity dates creates a reusable layer across channels and future internal use cases.
7. Human-agent complementarity supports adoption. The agent expands availability, prepares cases, and transfers context; humans retain judgment, empathy, negotiation, and exception handling.
8. Think globally, build rung by rung. Tools, use cases, and development methods are evolving at the same time. I therefore recommend an architectural vision from the outset, followed by a first real,

measurable scope that is expanded progressively. This is the method proposed here; other paths may be appropriate depending on context.

ABOUT SACHA OMON AI

Who publishes this White Paper, how the company works, and then the founder. The appendices - market overview, glossary, sources - begin afterward on their own page.

About SACHA OMON AI

From enterprise knowledge to multichannel action: web, WhatsApp, phone, email, and SMS. Designed and developed in Israel for the domestic and international markets.

SACHA OMON AI designs and operates an AI SDR agent that acts as soon as inbound demand appears: it understands, qualifies, provides the expected answer or document, proposes the next step, and hands off to the human team when necessary. This page explains what the company does, how it works, and the four choices that distinguish it from other market solutions.

WHAT WE DO

One agent, every channel, operated by our teams

The multichannel agentic AI SDR agent acts as soon as inbound demand appears across web, WhatsApp, phone, email, and SMS, with centralized context, 24/7 availability, and support for more than 40 languages. A conversation started on one channel continues on another without requiring the person to repeat themselves. The technical foundation is Sacha Engine, developed and operated by the company.

SDR means sales development representative: in a sales organization, this is the person who receives inbound inquiries, checks whether they fit the offer, filters out those that do not, and passes to sales the ones that deserve a meeting. That is the role the agent performs across all channels at once. The service target is a first action in under two seconds, at any hour - where conventional response time is still measured in hours, sometimes days. Chapter 01 shows what that delay costs.

2 s	24/7	40+	1	0
First-action objective	Always-on availability, all channels	Supported languages	One agent, one shared context	Commission on completed sales

Source: SACHA OMON AI • www.sacha-omon.ai

CHOICE NO. 1 · TURNKEY

The market sells tools to configure. We deliver a service that works.

Most AI-agent solutions are platforms. The company buys a license, then does everything itself: describe the offer, gather and sort knowledge, write the rules, connect the CRM, test, fix, and repeat. That time is almost never counted when the decision is made. It is always paid for later.

SACHA OMON AI chose the opposite approach. Understanding a business, analyzing its inbound demand, and translating its commercial reasoning into an agent is our job - not our customer's. We therefore take responsibility for the project end to end.

PROJECT STAGE	SELF-CONFIGURED PLATFORM	SACHA OMON AI
Analysis of the business and inbound requests	The customer	SACHA OMON AI
Knowledge base: collection, sorting, validation	The customer	SACHA OMON AI
Rules, guardrails, human-handoff scenarios	The customer	SACHA OMON AI
Integrations: CRM, calendar, catalog, support	The customer	SACHA OMON AI
Testing, corrections, and production launch	The customer	SACHA OMON AI
Monitoring, measurement, and improvements	The customer	SACHA OMON AI · dedicated project manager
Ownership of accounts, data, and created assets	The customer	The customer

Table 50: The customer retains ownership of everything. SACHA OMON AI carries the workload.

Many solutions ask the customer to do everything. We chose to do it for them. It is simpler, and it is faster.

CHOICE NO. 2 · MONTHLY OPTIMIZATION

An agent that is not reviewed every month will eventually get things wrong

This market moves faster than any one company can reasonably track. Models change, channels change, use cases change, rules change. An online course can be partly outdated a few weeks later. That is not poor organization or a lack of rigor; it is the pace of the market.

That is the lesson of more than thirty years in web development and digital acquisition: a company is stronger when it is surrounded by the right expertise. We apply that principle to AI. The agent is not delivered and abandoned - it is measured, corrected, and upgraded every month by our teams.

01	Measure Volumes handled, response quality, human-handoff rate, meetings and opportunities generated.
02	Analyze Questions without satisfactory answers, new objections, missing or outdated sources.
03	Correct Updates to the knowledge base, rules, scenarios, and integrations.
04	Upgrade Version upgrades, new models, new channels, and new capabilities, without customer intervention.

A dedicated project manager follows the account. The customer has no model to monitor, no version to install, and no technical vendors to coordinate.

CHOICE NO. 3 • SINGLE TENANT

Dedicated infrastructure for every customer, never shared

Two architectures coexist in enterprise software. In multi-tenant architecture, all of a vendor's customers share the same application and database; their data is separated by software rules. In single-tenant architecture, each customer has its own instance and its own database. SACHA OMON AI chose single tenant.

COMPARISON POINT	MULTI-TENANT	SINGLE-TENANT • OUR CHOICE
Database	Shared across the vendor's customers	Private PostgreSQL database dedicated to the customer
Data isolation	Logical, managed by the software	Physical, through a dedicated instance
Shared environment	Another customer's activity can affect yours	No shared tenancy
Configuration	Shared, with limited options	Adapted to the customer's rules and processes
Reversibility	Extraction dependent on the vendor	Identifiable, exportable perimeter
Compliance	Must be demonstrated at platform level	Must be demonstrated within a clearly defined perimeter

Table 51: Single-tenant architecture costs more to operate. It avoids confusing use of a service with ownership of the enterprise's AI asset.

This choice has a direct consequence for leadership: enterprise knowledge, conversations, rules, and assets created during the project remain inside a perimeter that belongs to the customer. The intelligence layer - the model - remains external and selectable, and receives only the information needed for the authorized task. Processing is aligned with GDPR requirements where applicable.

DESIGN CHOICE #4 • AGENTIC AI + RAG

RAG provides accuracy. Agentic AI provides action. You need both.

These two terms appear throughout the White Paper. In three lines, here is what they mean and why their combination is at the heart of our architecture.

BUILDING BLOCK	WHAT IT DOES	WHAT IS MISSING WITHOUT IT
LLM only	Understands language and generates a fluent response	It does not know your offers, prices, or rules: it can make things up
RAG	Retrieves the answer from the company's validated documents before responding	Without RAG, the answer may be plausible but is not verifiable
Agentic AI	Receives an objective, plans, uses tools, observes the outcome, and corrects course through completion	Without agentic AI, the system answers but does not act: no meeting, no document, no follow-up

Table 52: Three building blocks, three distinct roles. None replaces the other two.

WHY THE COMBINATION CHANGES THE OUTCOME

RAG gives the agent accurate, dated, attributable knowledge: it answers from company data, not from training memory.

Agentic AI gives that knowledge the ability to act: qualify, propose a time slot, send a document, follow up, hand off to sales with full context.

Together, they produce what a company actually expects: a request carried through into actions, from the first message to resolution - with traceability at every step.

Automation and agentic AI are often confused even though they address different needs. An automation system executes a sequence programmed in advance, identical at every trigger - useful for a stable, repeatable process. An agentic system chooses its path at each interaction based on the objective and what it observes - necessary whenever the request varies, as nearly all inbound sales requests do. The same distinction applies to rule-based chatbots, which answer within a closed scenario defined in advance (see Table 20, page 49). SACHA OMON AI belongs to a third category: agentic, not automation and not a chatbot.

THE FOUNDER

Salomon Bellaïche - thirty years of digital acquisition, now applied to agentic AI

Before digital, Salomon Bellaïche spent eight years in audiovisual production, publishing, and music. In 1996, he founded one of France's first digital agencies focused on digital acquisition and demand generation. Since then, he has worked with more than 3,000 companies, from small businesses to large enterprises, including Microsoft and L'Oréal.

The throughline has remained the same: turn marketing investment into real demand, understand intent, qualify the need, and move the relationship toward a commercial outcome. More than three years of hands-on work in call centers in Israel added direct observation of conversations, objections, transfers, and the moments when humans need to take over.

Today, that experience is applied to SACHA OMON AI. Its development represents nearly four years of work, and the company has been self-funded since its creation. Sacha Engine combines LLMs, RAG, and agentic orchestration to connect enterprise knowledge, tools, channels, and teams. SACHA OMON AI belongs to a third category: neither automation nor chatbot. Automation executes a predefined path; a chatbot converses; an agentic system understands a situation, retrieves the required knowledge, chooses among authorized actions, and acts within a controlled framework. Designed and developed in Israel, the solution targets both the Israeli and international markets, with an architecture designed for the United States, Europe, and Asia. The objective: understand every need expressed by a prospect, customer, or citizen, provide the right answer, and trigger the appropriate action to satisfy that need, 24/7, through AI.

Founder profile: [linkedin.com/in/salomon-sylvain-bellaïche-8187796](https://www.linkedin.com/in/salomon-sylvain-bellaïche-8187796) • Website: www.sacha-omon.ai • Demo: www.sacha-omon.com

GO FURTHER

Strategic assessment: analysis of your inbound requests, channels, and first-response time, with an estimate of opportunities currently lost, delayed, or insufficiently handled. -> www.sacha-omon.ai

Standalone requirements document: use the script on pages 89 to 92 in your own AI to scope your first use case and produce a first version of the document. No online simulator or scoping agent is required.

30-minute demonstration with opportunity assessment, no commitment. -> www.sacha-omon.com

Direct contact: sacha@sacha-omon.com

Founder's LinkedIn: [linkedin.com/in/salomon-sylvain-bellaïche-8187796](https://www.linkedin.com/in/salomon-sylvain-bellaïche-8187796)

© 2026 SACHA OMON AI. All rights reserved. Quotation is permitted with source attribution; reproduction requires written permission. All trademarks cited belong to their respective owners. Market data is reproduced for informational purposes and refers back to its original publishers; it has not been independently verified by SACHA OMON AI and remains the responsibility of its respective authors. This document is provided for informational purposes only: it does not constitute legal, financial, or investment advice, nor any contractual commitment by SACHA OMON AI.

WHITE PAPER APPENDICES

Model-provider overview, glossary, methodology, sources, and timeline - to verify, go further, and cite this White Paper accurately.

Appendix A - Overview of model providers and their engines

This White Paper is intentionally vendor-agnostic: it describes architectures, not products. This table lists the principal language-model providers cited in the White Paper, their country of origin, and the URL of their official documentation so readers can verify the current state of each offering rather than rely on this document.

PUBLISHER	ORIGIN	MODEL ENGINES AND FAMILIES	OFFICIAL DOCUMENTATION
OpenAI	United States	GPT, ChatGPT	openai.com/index · platform.openai.com/docs
Anthropic	United States	Claude : Opus, Sonnet, Haiku	anthropic.com/news · docs.anthropic.com
Google and Google DeepMind	United States	Gemini : Pro, Flash	blog.google · ai.google.dev
Microsoft	United States	Copilot, Phi, Azure AI Foundry	microsoft.com/copilot · learn.microsoft.com
Meta	United States	Llama, Meta AI	ai.meta.com/llama
xAI	United States	Grok	x.ai/news · docs.x.ai
Amazon Web Services	United States	Nova, plateforme Bedrock	aws.amazon.com/bedrock
Perplexity	United States	Generative search engine	perplexity.ai
Mistral AI	France	Mistral, Magistral, Le Chat	mistral.ai/news · docs.mistral.ai
Cohere	Canada	Command, Embed, Rerank	cohere.com · docs.cohere.com
DeepSeek	China	DeepSeek V, DeepSeek R	api-docs.deepseek.com
Alibaba Cloud	China	Qwen	qwen.ai · alibabacloud.com
Moonshot AI	China	Kimi	moonshot.ai · platform.moonshot.ai
Zhipu AI	China	GLM	z.ai · bigmodel.cn
ByteDance	China	Doubao	volcengine.com
EPFL, ETH Zurich, and CSCS	Switzerland	Apertus	apertus.ai · huggingface.co/swiss-ai

Table 53: Non-exhaustive overview as of August 19, 2026. Versions and pricing change constantly; only each provider's official documentation is authoritative.

None of these providers participated in writing, reviewing, or funding this White Paper. The architectures described here - agentic AI and RAG - are independent of the provider selected; that is precisely the value of the multi-model architecture described in Chapter 7.

Appendix B - Glossary: terms and definitions

TERM	DEFINITION
A2A (Agent2Agent)	Open protocol published by Google in April 2025 and now governed by the Linux Foundation, allowing AI agents to discover one another and coordinate actions.
ACP (Agentic Commerce Protocol)	Open standard co-developed by OpenAI and Stripe that structures exchanges among an AI agent, a buyer, and a merchant to complete a purchase.
Agentic agent	System that receives an objective, plans, calls tools, observes, corrects, and drives the task to a terminal state.
Agentic RAG	Architecture combining agentic reasoning and document retrieval: the agent decides what to search for, where, how many times, and evaluates whether the evidence is sufficient.
AP2 (Agent Payments Protocol)	A2A extension published by Google that represents each payment executed by an agent as cryptographically signed mandates proving user authorization.
Apertus (EPFL, ETH Zurich, CSCS)	Multilingual open-weight model family developed by a Swiss academic consortium (EPFL, ETH Zurich, and the Swiss National Supercomputing Centre). Documentation: apertus.ai and huggingface.co/swiss-ai .
Tool call	Structured invocation of an external function by the agent: search a document, read a calendar, or create a CRM opportunity.
Multi-model architecture	Architecture capable of using multiple providers or model families without locking company knowledge and processes into a single engine.
Knowledge base	Governed set of validated, attributed, dated sources accessible under explicit permissions, used by one or more AI systems.
Conversational channel	Interface through which the user interacts with the agent: web, WhatsApp, phone, email, SMS, or another messaging channel.
ChatGPT and GPT (OpenAI)	ChatGPT is OpenAI's consumer application (United States); GPT refers to the model family that powers it. Documentation: openai.com/index and platform.openai.com/docs .
Chunking	Segmenting documents into coherent chunks before vectorization.

TERM	DEFINITION
Claude (Anthropic)	Language-model family developed by Anthropic (United States), available in Opus, Sonnet, and Haiku variants. Documentation: anthropic.com/news and docs.anthropic.com .
Copilot (Microsoft)	Family of AI assistants integrated into Microsoft products (United States), backed by the Azure AI Foundry platform. Documentation: microsoft.com/copilot and learn.microsoft.com/azure/ai-foundry .
Intelligence layer (LLM)	Component responsible for understanding, reasoning, planning, synthesizing, and formulating; it does not replace the knowledge base, memory, or permissions.
DeepSeek	Chinese open-weight model provider known for the V and R families and for its performance-to-cost ratio. Documentation: api-docs.deepseek.com .
Doubao (ByteDance)	Model family and consumer assistant developed by ByteDance (China), widely used in China's domestic market. Documentation: volcengine.com .
Embedding	Numerical representation of text enabling semantic comparison.
Escalation (human handoff)	Controlled handoff from the agent to a person, including identity, history, consulted sources, and the recommended next action.
Faithfulness	Measure of how well the generated answer aligns with the sources provided.
Garde-fou	Deterministic control that constrains agent behavior independently of the model.
Gemini (Google)	Multimodal model family from Google and Google DeepMind (United States), available in Pro and Flash variants. Documentation: blog.google and ai.google.dev .
GLM (Zhipu AI)	Open-weight model family developed by Zhipu AI (China). Documentation: bigmodel.cn and z.ai .
GraphRAG	RAG variant that uses a knowledge graph to connect entities.
Grok (xAI)	Model family developed by xAI (United States), integrated with the X platform and available via API. Documentation: x.ai/news and docs.x.ai .
Hallucination	Generation of plausible but false information in the absence of a source.
Kimi (Moonshot AI)	Long-context model family developed by Moonshot AI (China). Documentation: moonshot.ai and platform.moonshot.ai .

TERM	DEFINITION
Llama (Meta)	Open-weight model family released by Meta (United States), widely used for self-hosted deployments. Documentation: ai.meta.com/llama .
LLM	Large Language Model, the engine for understanding and generating text.
MCP (Model Context Protocol)	Open standard for connecting AI systems and tools, governed since December 2025 by the Agentic AI Foundation.
Persistent memory	Retention of context and history beyond a single session.
Mistral AI	French language-model provider offering both open-weight and proprietary models, the Le Chat assistant, and an agent API. A leading European foundation-model company. Documentation: mistral.ai/news and docs.mistral.ai .
Mono-locataire	Architecture in which each customer has dedicated infrastructure and a dedicated database.
Multi-sauts (multi-hop)	Reasoning that requires multiple successive searches, each informed by the previous one.
Nova and Bedrock (Amazon)	Nova is Amazon's model family (United States); Bedrock is the AWS platform providing access to multiple model families. Documentation: aws.amazon.com/bedrock .
Opt-in (prior consent)	A regime in which commercial outreach is lawful only after the person being contacted has given explicit prior consent, as opposed to opt-out, where outreach is permitted unless the person has objected in advance.
Orchestration	Layer that coordinates reasoning, rules, memory, document retrieval, tools, and process termination.
Perplexity	Generative AI search engine (United States) that produces answers with citations. Documentation: perplexity.ai .
First useful response	The first response that genuinely advances the request; distinct from an acknowledgment with no operational value.
Qwen (Alibaba Cloud)	Open-weight model family developed by Alibaba Cloud (China), among the most widely adopted globally for self-hosting. Documentation: qwen.ai and alibabacloud.com .
RAG	Retrieval-Augmented Generation: generation of answers grounded in a document base queried at the time of the question.
Dedicated RAG	Knowledge base and retrieval pipeline isolated for one company or entity, with no data mixing across customers.

TERM	DEFINITION
Reranking	Fine-grained reranking of candidate passages by a specialized model after initial retrieval.
Model routing	Selection by the orchestrator of the model best suited to a task based on language, latency, reasoning depth, cost, and data constraints.
Authoritative source	Document, system, or editorial owner recognized as the reference when multiple versions of information exist.
Speed-to-lead	Time between arrival of an inbound request and the first useful response.
Human oversight	Human validation or intervention mechanism built into the agent's journey.
Token	Elementary unit used by an LLM to process information. APIs generally charge for input and output tokens; volume depends on the context transmitted, reasoning steps, and generated response. Tokens are an important component of LLM cost, but not the only cost in an agentic system.
Agentic work unit	Discrete task completed by an agent: decision made, record updated, process triggered.
Agentic web	Emerging evolution of the Internet in which AI agents discover one another, authenticate, and exchange directly, beyond simple human navigation on websites.

Table 54: Glossary of terms used in this White Paper, in a single alphabetical sequence.

Appendix C - Methodology

This White Paper relies exclusively on public sources consulted and verified through August 19, 2026. Four categories of sources were used.

1. Analyst firms: Gartner, McKinsey, IDC, Deloitte, Forrester, and specialized market-research firms, for adoption, spending, and forecast data.
2. Official vendor communications: product announcements, earnings communications, and case-study pages from Amazon, AWS, Salesforce, Microsoft, Google, OpenAI, and Anthropic.
3. Academic research: arXiv publications and peer-reviewed papers on RAG, agentic RAG, and faithfulness evaluation, selected for quantified and reproducible results.
4. Leading business and technology press: TechCrunch, CNBC, Axios, Fortune, Forbes, GeekWire, for dated corporate facts.

Three treatment principles were applied. Figures originating from vendors are identified as such because commercial communications require cautious interpretation. Differences between analyst firms are disclosed rather than arbitrated, since they often measure different scopes. Unverifiable data, or data attributed to organizations with no documented existence, was excluded even when widely repeated.

Appendix D - Clickable sources and articles

Analyst firms, research, and press

ORGANIZATION	REFERENCE	ACCESS
McKinsey & Company	The state of AI in 2025: Agents, innovation, and transformation (Nov. 2025)	www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/november%202025/the-state-of-ai-2025-agents-innovation_cmyk-v1.pdf
McKinsey & Company	State of AI trust in 2026: Shifting to the agentic era	mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era
McKinsey & Company	Reinventing marketing workflows with agentic AI	mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/reinventing-marketing-workflows-with-agentic-ai
Project NANDA / MIT Media Lab	Aditya Challapally, Chris Pease, Ramesh Raskar, Pradyumna Chari — The GenAI Divide: State of AI in Business 2025 — Preliminary Findings (July 2025)	www.media.mit.edu/groups/nanda/overview/
MIT Media Lab	Project NANDA : the Internet of AI Agents, NANDA Index, AgentFacts	nanda.media.mit.edu
Gartner	2026 agentic spending forecasts, application adoption, 2027 cancellation rate	gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025 ; gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027
Srinivasan et al. (IIT Madras)	Enhancing Financial RAG with Agentic AI and Multi-HyDE	arxiv.org/abs/2509.16369
Multiple authors	Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG	arxiv.org/abs/2501.09136
Multiple authors	An Empirical Study of Multi-Agent RAG (MARAU)	arxiv.org/abs/2507.11272
Multiple authors	Agentic Hybrid RAG for Evidence-Grounded Analysis	arxiv.org/abs/2606.10381
Multiple authors	Agent-Orchestrated Adaptive RAG: A Comparative Study	arxiv.org/abs/2606.05658

ORGANIZATION	REFERENCE	ACCESS
MDPI	Evaluating Faithfulness in Agentic RAG Systems, Big Data Cogn. Comput. 9(12), 309	mdpi.com/2504-2289/9/12/309
Mordor Intelligence	Retrieval Augmented Generation Market	mordorintelligence.com/industry-reports/retrieval-augmented-generation-market
MarketsandMarkets	Retrieval-augmented Generation (RAG) Market, Global Forecast to 2030	marketsandmarkets.com/Market-Reports/retrieval-augmented-generation-rag-market-135976317.html
Next Move Strategy Consulting	RAG Market Outlook 2035	nextmsc.com/report/retrieval-augmented-generation-rag-market-ic3918
Fortune Business Insights / Precedence Research / Deloitte	Agentic AI market estimates	fortunebusinessinsights.com/agentic-ai-market-114233 (Fortune Business Insights); precedenceresearch.com/agentic-ai-market (Precedence Research); Deloitte: no dedicated public report page identified
The Business Research Company	AI Sales Development Representative (SDR) Global Market Report (July 2026)	thebusinessresearchcompany.com/report/artificial-intelligence-ai-sales-development-representative-sdr-global-market-report
Kantar for Meta	The State of Business Messaging (11,056 respondents, 22 markets)	whatsappbusiness.com/resources/resource-library/state-of-business-messaging/
Sacra	Sierra and Decagon revenue analyses (2026)	sacra.com/research/sierra-vs-decagon/
Juniper Research	Global business messaging traffic 2025-2030	juniperresearch.com/press/conversational-use-cases-fuel-global-messaging-boom/
EMARKETER	Meta to Surpass Google in Digital Ad Revenues for First Time Ever (Apr. 2026)	emarketer.com/press-releases/meta-to-surpass-google-in-digital-ad-revenues-for-first-time-ever/
ResearchAndMarkets	United States Digital Ad Spend Market Size & Forecast (Feb. 2026)	researchandmarkets.com/reports/6217876/united-states-digital-ad-spend-market-size-and
HubSpot	Share of phone solicitations received outside standard business hours	hubspot.com
Foureyes / Ringlead	Internet Lead Response Statistics for Dealerships: study of 22,500 dealerships, 2026	ringlead.ca/blog/speed-to-lead/internet-lead-response-statistics-2026/
Reuters	Verizon says Google AI for customer service agents has led to sales jump (Apr. 2025)	reuters.com/business/media-telecom/verizon-says-google-ai-customer-service-agents-has-led-sales-jump-2025-04-09/

ORGANIZATION	REFERENCE	ACCESS
Reuters	Teleperformance shares plunge on AI disruption concerns (Feb. 28, 2024)	reuters.com/technology/teleperformance-shares-plunge-ai-disruption-concerns-2024-02-28/
Euronext Paris	TP SE: market price and history, close on Aug. 7, 2026	live.euronext.com/en/product/equities/FR0000051807-XPAR
TP	Investor information, H1 2026 results, and TP.AI positioning	tp.com/media/1yxhdsni/tp-press-release-h1-2026-results.pdf
Financial Times	Hedge funds bet against call-centre groups as AI disruption accelerates (June 2026)	ft.com
TechCrunch / CNBC / Axios	Prometheus: \$12 billion raise at a \$41 billion valuation (June 2026)	techcrunch.com/2026/06/11/jeff-bezoss-prometheus-raises-12b-to-build-an-artificial-general-engineer-for-the-physical-world/ ; cnbc.com/2026/06/11/project-prometheus-bezos-bajaj-live-updates.html ; axios.com/2026/06/11/prometheus-bezos-industrial-ai
Marc Benioff	The Logan Bartlett Show, episode 149, Aug. 29, 2025	podcasts.apple.com/us/podcast/ep-149-marc-benioff-ceo-salesforce-predicts-half-of/id1606770839?i=1000724017332
Salesforce / SEC	Form 10-Q, quarter ended April 30, 2026 - acquisition of Qualified.com, Inc.; fair value of consideration: \$1.2 billion, primarily \$1.1 billion in cash	www.sec.gov/Archives/edgar/data/1108524/000110852426000127/0001108524-26-000127-index.htm
Digital Applied	AI SDR Real Performance: 100K Email Analysis 2026 (April 26, 2026)	digitalapplied.com/blog/ai-sdr-real-performance-100k-email-analysis-2026
Drag	AI Support Agent Resolution Rates: What Vendors Claim vs What Production Shows (Aug. 15, 2026)	dragapp.com/blog/ai-support-agent-resolution-rates
Israel Innovation Authority	Israeli High-Tech Status Report 2026	innovationisrael.org.il/en/press_release/israeli-high-tech-report-2026/
Startup Nation Central	Israel's World-Class AI Powerhouse: Leading Through Applied Innovation	finder.startupnationcentral.org/reports/ai-report-2024

Table 55: Analysis, research, and press sources.

Vendors, platforms, and official documentation

ORGANIZATION	REFERENCE	ACCESS
Salesforce	Agentforce Metrics; FY2026 results; help-site case study	salesforce.com/agentforce/metrics
Salesforce	Agentforce launch at Dreamforce (Sept. 17, 2024)	salesforce.com/news/stories/dreamforce-24-recap/

ORGANIZATION	REFERENCE	ACCESS
Salesforce	Agentic Enterprise Index H1 2025 (Sept. 17, 2025) and 2025-2026 (Aug. 7, 2026)	salesforce.com/news/stories/agentic-enterprise-index-insights-h1-2025/ ; salesforce.com/news/stories/agentic-enterprise-index-insights-2026/
Salesforce	Signs Definitive Agreement to Acquire Qualified (Dec. 17, 2025)	salesforce.com/news/stories/salesforce-signs-definitive-agreement-to-acquire-qualified/
Salesforce	State of Sales 2026, 7th edition (Feb. 2026)	salesforce.com/sales/state-of-sales
Qualified	Piper case studies, including Demandbase	qualified.com/customers
IBM Think	Study on how AI SDR agents are redefining sales (Apr. 2026)	ibm.com/think/topics/ai-sdr
Amazon / AWS	AWS Summit New York 2026: agents, managed Knowledge Bases, AWS Context, Web Search	aboutamazon.com/news/aws/aws-summit-nyc-2026-ai-agents
Amazon	Q2 2026 results: Andy Jassy comments on AgentCore	aboutamazon.com/news/company-news/amazon-ceo-andy-jassy-aws-revenue-growth-q2-2026-earnings
AWS	Agents and Knowledge Bases for Amazon Bedrock preview (Sept. 13, 2023) and availability (Nov. 28, 2023)	aws.amazon.com/blogs/aws/preview-connect-foundation-models-to-your-company-data-sources-with-agents-for-amazon-bedrock/ ; aws.amazon.com/about-aws/whats-new/2023/11/amazon-bedrock-knowledge-bases/
AWS	Amazon Bedrock AgentCore: preview (July 16, 2025), general availability (Oct. 13, 2025)	aws.amazon.com/blogs/aws/introducing-amazon-bedrock-agentcore-securely-deploy-and-operate-ai-agents-at-any-scale ; aws.amazon.com/about-aws/whats-new/2025/10/amazon-bedrock-agentcore-available
AWS	AgentCore Policy (Mar. 3, 2026), Evaluations (Mar. 31, 2026), Policy with Guardrails (June 17, 2026)	aws.amazon.com/about-aws/whats-new/2026/03/policy-amazon-bedrock-agentcore-generally-available/ ; aws.amazon.com/about-aws/whats-new/2026/03/agentcore-evaluations-generally-available/ ; aws.amazon.com/about-aws/whats-new/2026/06/amazon-bedrock-agentcore-policy-guardrails-generally-available/

ORGANIZATION	REFERENCE	ACCESS
AWS	Agentic RAG in Amazon Q Business (Aug. 14, 2025); Hybrid RAG (Apr. 6, 2026); Managed Knowledge Base (July 16, 2026)	aws.amazon.com/about-aws/whats-new/2025/08/qbusiness-launches-agentic-rag/ ; aws.amazon.com/blogs/machine-learning/build-enterprise-search-for-agents-with-amazon-bedrock-managed-knowledge-base/
AWS	Bedrock Agents Classic maintenance mode (July 30, 2026)	docs.aws.amazon.com/bedrock
Anthropic	Introducing Claude (Mar. 14, 2023); Claude 3 (Mar. 4, 2024); Claude 3.5 Sonnet (June 21, 2024)	anthropic.com/news/introducing-claude ; anthropic.com/news/claude-3-family ; anthropic.com/news/claude-3-5-sonnet
Anthropic	Contextual Retrieval (Sept. 19, 2024); Building Effective Agents (Dec. 19, 2024)	anthropic.com/engineering/contextual-retrieval ; anthropic.com/engineering/building-effective-agents
Anthropic	Introducing the Model Context Protocol (Nov. 25, 2024); MCP donation and Agentic AI Foundation (Dec. 9, 2025)	anthropic.com/news/model-context-protocol ; anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation
Anthropic	Claude 3.7 Sonnet (Feb. 24, 2025); Claude 4 (May 22, 2025); multi-agent research system (June 13, 2025)	anthropic.com/news/claude-3-7-sonnet ; anthropic.com/news/claude-4 ; anthropic.com/engineering/multi-agent-research-system
Anthropic	Effective context engineering for AI agents (September 29, 2025) ; Anthropic Labs (13 Jan. 2026)	anthropic.com/engineering/effective-context-engineering-for-ai-agents ; anthropic.com/news/introducing-anthropic-labs
Anthropic	Claude Opus 4.6 (Feb. 5, 2026); Sonnet 4.6 (Feb. 17, 2026); Fable 5 and Mythos 5 (June 9, 2026); Sonnet 5 (June 30, 2026); Opus 5 (July 24, 2026)	anthropic.com/news/claude-opus-4-6 ; anthropic.com/news/claude-sonnet-4-6 ; anthropic.com/news/claude-fable-5-mythos-5 ; anthropic.com/news/claude-sonnet-5 ; anthropic.com/news/claude-opus-5
OpenAI	Improving Language Understanding (June 11, 2018); GPT-2 (Feb. 14, 2019); GPT-3 (2020); OpenAI API (June 11, 2020)	openai.com/index/language-unsupervised/ ; openai.com/index/better-language-models/ ; openai.com/index/language-models-are-few-shot-learners/ ; openai.com/index/openai-api/

ORGANIZATION	REFERENCE	ACCESS
OpenAI	WebGPT (Dec. 16, 2021); InstructGPT (Jan. 27, 2022); ChatGPT (Nov. 30, 2022); GPT-4 (Mar. 14, 2023)	openai.com/index/webgpt/ ; openai.com/index/instruction-following/ ; openai.com/index/chatgpt/ ; openai.com/index/gpt-4-research/
OpenAI	Function calling (June 13, 2023); GPT-4o (May 13, 2024); o1-preview (Sept. 12, 2024); o3 and o4-mini (Apr. 16, 2025)	openai.com/index/function-calling-and-other-api-updates/ ; openai.com/index/hello-gpt-4o/ ; openai.com/index/introducing-openai-o1-preview/ ; openai.com/index/introducing-o3-and-o4-mini/
OpenAI	New tools for building agents (Mar. 11, 2025); gpt-oss (Aug. 5, 2025); GPT-5 (Aug. 7, 2025); AgentKit (Oct. 6, 2025)	openai.com/index/new-tools-for-building-agents/ ; openai.com/index/introducing-gpt-oss/ ; openai.com/index/introducing-gpt-5/ ; openai.com/index/introducing-agentkit/
OpenAI	Computer environment for the Responses API (Mar. 11, 2026); Agents SDK (Apr. 15, 2026); Presence (July 22, 2026)	openai.com/index/equip-responses-api-computer-environment/ ; openai.com/index/the-next-evolution-of-the-agents-sdk/ ; openai.com/index/introducing-openai-presence/
OpenAI	GPT-5.4 (Mar. 5, 2026); GPT-5.5 (Apr. 23, 2026); GPT-5.6 Sol preview (June 26, 2026); GPT-5.6 (July 9, 2026); Sol and Luna in ChatGPT (Aug. 6, 2026)	openai.com/index/introducing-gpt-5-4/ ; openai.com/index/introducing-gpt-5-5/ ; openai.com/index/previewing-gpt-5-6-sol/ ; openai.com/index/gpt-5-6/ ; openai.com/index/improving-gpt-5-6-sol-in-chatgpt/
OpenAI / Stripe	Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol (Sept. 2025-Feb. 2026)	openai.com/index/buy-it-in-chatgpt
Google Research	Attention Is All You Need (2017) ; BERT (2018)	arxiv.org/abs/1706.03762 ; arxiv.org/abs/1810.04805
Google	Gemini 1.0 (Dec. 6, 2023); Gemini 2.0 (Dec. 11, 2024); Gemini 2.5 (Mar. 25, 2025); Gemini 3 (Nov. 18, 2025)	blog.google/technology/ai/google-gemini-ai/ ; blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/ ; blog.google/innovation-and-ai/models-and-research/google-deepmind-gemini-model-thinking-updates-march-2025/ ; blog.google/products/gemini/gemini-3/

ORGANIZATION	REFERENCE	ACCESS
Google	Gemini 3.1 Pro (Feb. 19, 2026); Gemini 3.5 Flash (May 19, 2026); Gemini 3.6 Flash (July 21, 2026)	blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/ ; blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/ ; blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/
Google	Agent2Agent Protocol (April 9, 2025) ; Agent Payments Protocol AP2 (September 16, 2025) ; Agentic Resource Discovery (June 17, 2026)	developers.googleblog.com/en/a2a-a-new-era-of-agent-interopability/ ; cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol ; developers.googleblog.com/announcing-the-agentic-resource-discovery-specification/
Google	Gemini Embedding (July 30, 2025); Gemini Embedding 2 (March and Apr. 30, 2026)	developers.googleblog.com/gemini-embedding-powering-rag-context-engineering/ ; blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-embedding-2/
Google Cloud	Vertex AI RAG Engine (15 Jan. 2025) ; Gemini Enterprise Agent Platform (April 22, 2026)	cloud.google.com/blog/products/ai-machine-learning/introducing-vertex-ai-rag-engine ; cloud.google.com/blog/products/ai-machine-learning/introducing-gemini-enterprise-agent-platform
Google	Managed Agents in Gemini API (May 19, 2026); budget controls and tool audits (July 28, 2026)	blog.google/innovation-and-ai/technology/developers-tools/managed-agents-gemini-api/ ; blog.google/innovation-and-ai/technology/developers-tools/expanding-managed-agents-gemini-api-3-6-flash-hooks/
Google and partners	Agent Plugins 1.0.0 (Aug. 6, 2026)	developers.googleblog.com/agent-plugins-package-your-skills-tools-and-more/
Google	Announces Google I/O 2026 : agent d'achat personnel, Business Agent (May 2026)	blog.google/innovation-and-ai/technology/ai
Meta AI	Dense Passage Retrieval (2020); LLaMA (Feb. 24, 2023); Llama 2 (July 18, 2023); Llama 3 (Apr. 18, 2024); Llama 3.1 (July 23, 2024); Llama 4 (Apr. 5, 2025)	research.facebook.com/publications/dense-passage-retrieval-for-open-domain-question-answering/ ; ai.meta.com/blog/large-language-model-llama-meta-ai/ ; ai.meta.com/blog/llama-2/ ; ai.meta.com/blog/meta-llama-3/ ; ai.meta.com/blog/meta-llama-3-1/ ; ai.meta.com/blog/llama-4-multimodal-intelligence/

ORGANIZATION	REFERENCE	ACCESS
Microsoft Research	AutoGen (Sept. 25, 2023); GraphRAG (Apr. 2024); GraphRAG 1.0 (Dec. 16, 2024)	microsoft.com/en-us/research/blog/autogen-enabling-next-generation-large-language-model-applications/ ; microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/
Mistral AI	Mistral 7B (Sept. 27, 2023); Agents API (May 27, 2025); Mistral 3 (Dec. 2, 2025); Medium 3.5 and Work mode (May 22, 2026)	mistral.ai/news/announcing-mistral-7b/ ; mistral.ai/news/agents-api
xAI	Grok-2 (Aug. 13, 2024); Grok 3 (Feb. 19, 2025); Grok 4 (July 9, 2025); Grok Models and Agent Tools	x.ai/news , docs.x.ai
DeepSeek	DeepSeek-R1 (Jan. 20, 2025); V3.2 (Dec. 1, 2025); V4 Preview (Apr. 24, 2026)	api-docs.deepseek.com/news/news250120
Model Context Protocol	Specification 2026-07-28 (July 28, 2026)	blog.modelcontextprotocol.io
Self-RAG	Learning to Retrieve, Generate, and Critique through Self-Reflection (17 Oct. 2023)	arxiv.org/abs/2310.11511
CRAG	Corrective Retrieval-Augmented Generation (29 Jan. 2024)	arxiv.org/abs/2401.15884
Lewis et al.	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2020)	arxiv.org/abs/2005.11401
Yao et al.	ReAct: Synergizing Reasoning and Acting in Language Models (2022)	arxiv.org/abs/2210.03629
Schick et al.	Toolformer: Language Models Can Teach Themselves to Use Tools (2023)	arxiv.org/abs/2302.04761
MRKL Systems	A Modular, Neuro-Symbolic Architecture for Autonomous Agents (2022)	arxiv.org/abs/2205.00445

Table 56: Vendor, platform, and official-documentation sources.

Legal and regulatory sources

JURISDICTION	REFERENCE	ACCESS
France	French Law No. 2025-594 of June 30, 2025; Decree No. 2026-662 of July 23, 2026; Article L223-1 of the French Consumer Code	legifrance.gouv.fr/jorf/id/JORFTEXT000051824277 ; legifrance.gouv.fr/jorf/id/JORFTEXT00054483939 ; legifrance.gouv.fr/codes/article_lc/LEGIARTI000051830285/2026-08-11
France	DGCCRF: telephone solicitation rules applicable to businesses	economie.gouv.fr/dgccrf

JURISDICTION	REFERENCE	ACCESS
France	CNIL: commercial prospecting by email, SMS, MMS, and automated calling systems	cnil.fr/fr/la-prospection-commerciale-par-courrier-electronique-sms-mms-et-automate-dappel
European Union	GDPR, Regulation (EU) 2016/679; AI Act: implementation timeline and FAQs	eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32016R0679 ; digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act
European Commission	Withdrawal of the proposed ePrivacy Regulation (July 16, 2025)	Official Journal of the European Union, C/2025/5423
United States	FCC: AI-generated voices in robocalls (Feb. 8, 2024)	fcc.gov/document/fcc-confirms-tcpa-applies-ai-technologies-generate-human-voices
United States	Insurance Marketing Coalition v. FCC, U.S. Court of Appeals for the Eleventh Circuit, Jan. 24, 2025	public court decision
United States	Executive Orders 14179 and 14365; NIST AI Risk Management Framework	federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence ; federalregister.gov/documents/2025/12/16/2025-23092/ensuring-a-national-policy-framework-for-artificial-intelligence ; nist.gov/itl/ai-risk-management-framework
Texas / Colorado	HB 149 (Texas Responsible AI Governance Act) ; SB26-189 (Automated Decision-Making Technology)	capitol.texas.gov ; leg.colorado.gov
Israel	Responsible AI use guidance; Council of Europe Framework Convention on AI	harhayeda.gov.il/data-and-ai/policy/responsible-use-new/ ; gov.il
Singapore / Japan	Model AI Governance Framework for Agentic AI ; AI Guidelines for Business v1.2	imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf ; meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_12.pdf
China	Interim Administrative Measures for Generative AI Services	cac.gov.cn/2023-07/13/c_1690898327029107.htm
OWASP GenAI Security Project	LLM08:2025 : Vector and Embedding Weaknesses	genai.owasp.org
OpenAI	Business Data Privacy, Security and Compliance	openai.com/business-data/

Table 57: Legal and regulatory sources.

Appendix E - Clickable sources for the strategic opening

The facts and figures cited in the strategic opening rely on primary and official sources; any projection or interpretation reflects Salomon Bellaïche's editorial position. Prometheus is cross-checked against two leading media outlets because no equivalent institutional publication was available.

- [1] Zendesk CX Trends 2026 • 2026 (zendesk.com/blog/zendesk-insights/trends/cx-trends-2026-contextual-intelligence/)**
Signal on accelerating response-time expectations, always-on availability, and multimodal continuity.
- [2] Google I/O 2026 — Welcome to the agentic Gemini era • May 19, 2026 (blog.google/innovation-and-ai/sundar-pichai-io-2026/)**
Google presents Gemini Spark as a personal agent that acts on the user's behalf, under the user's direction.
- [3] Google — The Gemini app becomes more agentic • May 19, 2026 (blog.google/innovation-and-ai/products/gemini-app/next-evolution-gemini-app/)**
Gemini Spark is described as an always-on personal agent that works in the background and uses Google tools.
- [4] Meta - The Future is for Everyone • August 10, 2026 (about.fb.com/news/2026/08/the-future-is-for-everyone/)**
Mark Zuckerberg presents a vision of personal superintelligence placed in individuals' hands.
- [5] Meta — Meta Business Agent • June 3, 2026 (about.fb.com/news/2026/06/meta-business-agent/)**
Meta reports more than one million businesses using Business Agent and more than one billion active daily threads with businesses across WhatsApp, Messenger, and Instagram.
- [6] Meta — WhatsApp message translations • September 23, 2025 (about.fb.com/news/2025/09/introducing-message-translations-whatsapp/)**
Meta reports more than 3 billion WhatsApp users across more than 180 countries.
- [7] Microsoft — Human-led, agent-operated • April 16, 2026 (microsoft.com/insidetrack/blog/becoming-a-frontier-firm-a-guide-for-deploying-ai-agents-based-on-our-experience-at-microsoft/)**
Microsoft describes a model of work in which AI operates but humans lead.
- [8] Microsoft 365 Copilot, human agency • May 5, 2026 (microsoft.com/en-us/microsoft-365/blog/2026/05/05/microsoft-365-copilot-human-agency-and-the-opportunity-for-every-organization/)**
Microsoft frames it as "human-led, agent-operated, outcome-driven."
- [9] OpenAI — ChatGPT Work • July 9, 2026 (openai.com/index/chatgpt-for-your-most-ambitious-work/)**
ChatGPT Work brings information, apps, and files together to carry out long, multi-step tasks.
- [10] OpenAI — Workspace agents • April 22, 2026 (openai.com/index/introducing-workspace-agents-in-chatgpt/)**
OpenAI presents work agents that execute complex processes in the cloud.
- [11] OpenAI — Operator • 23 Jan. 2025 (openai.com/index/introducing-operator/)**
Operator is presented as an agent capable of using the web to carry out a task independently.
- [12] xAI — Grok Voice Agent Builder • July 2026 (x.ai/news/grok-voice-agent-builder)**
xAI offers a no-code builder for production voice agents.
- [13] xAI — Voice API • 2026 (x.ai/api/voice)**
Grok's Voice API supports real-time conversation, tools, and search.

[14] xAI — Connect Grok to Hermes Agent • May 15, 2026 (x.ai/news/grok-hermes)

Grok can be used in Hermes Agent, Nous Research's open-source persistent agent, which creates long-term memory across sessions and can connect to WhatsApp, Discord, Telegram, Signal, and other messaging services.

[15] xAI — Agent Tools API • 19 Nov. 2025 (x.ai/news/grok-4-1-fast)

xAI describes production agents specialized in tool use and agentic search.

[16] Apple — Siri AI • June 8, 2026 (apple.com/newsroom/2026/06/apple-introduces-siri-ai-a-profoundly-more-capable-and-personal-assistant/)

Siri AI uses personal context across Messages, email, Photos, and other apps and can act within apps.

[17] Amazon — Andy Jassy on generative AI • June 17, 2025 (aboutamazon.com/news/company-news/amazon-ceo-andy-jassy-on-generative-ai)

Andy Jassy anticipates billions of agents across companies and professions.

[18] NVIDIA — Q1 FY2027 results • May 20, 2026 (nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2027)

Jensen Huang says agentic AI is already producing work and value across enterprises and industries.

[19] Prometheus / Jeff Bezos — TechCrunch • June 11, 2026 (techcrunch.com/2026/06/11/jeff-bezoss-prometheus-raises-12b-to-build-an-artificial-general-engineer-for-the-physical-world/)

Prometheus, co-founded by Jeff Bezos and Vik Bajaj, announces \$12 billion in funding at an approximately \$41 billion valuation for AI focused on engineering and the physical world.

[20] Prometheus / Jeff Bezos — Financial Times • June 11, 2026

Bezos describes Prometheus as a project intended to transform engineering and manufacturing; this source cross-checks the funding and positioning.

[21] Otto — Hardware • 2026 (myotto.ai/hardware)

Otto One is presented as a computer dedicated to agents: credentials, files, and tasks remain local while inference still runs in the cloud.

[22] Otto — Presale • 2026 (myotto.ai/presale)

Otto One is offered starting at \$599, preloaded with OpenClaw; shipping is announced for January 2027.

[23] Otto — Contact • 2026 (myotto.ai/contact)

The Tokenry, Inc., the company behind Otto, says it is based in New York.

[24] Reid Hoffman & Greg Beato — Superagency • 2025 (simonandschuster.com/books/Superagency/Reid-Hoffman/9798893310108)

Hoffman develops the idea of AI augmenting human capacity for action.

[25] Mustafa Suleyman & Michael Bhaskar — The Coming Wave • 2023 (penguinrandomhouse.com/books/722674/the-coming-wave-by-mustafa-suleyman-with-michael-bhaskar/)

The book examines the diffusion of powerful systems and the problem of control.

[26] Kissinger, Schmidt & Mundie — Genesis • 2024 (hachettebookgroup.com/titles/henry-a-kissinger/genesis/9780316581295)

The book explores the consequences of AI gaining autonomy and the capacity to act.

[27] Légifrance - Article L223-1 of the French Consumer Code • August 11, 2026 (legifrance.gouv.fr/codes/article_lc/LEGIARTI000051830285/2026-08-11)

In France, B2C telephone solicitation shifts to prior consent, which the business must be able to prove.

[28] Légifrance - Decree No. 2026-662 · July 23, 2026

(legifrance.gouv.fr/jorf/id/JORFTEXT000054483939)

For certain sectors, the decree specifies the conditions for calling back after an explicit request for information and the five-business-day period.

[29] CNIL - Electronic marketing · 2026 (cnil.fr/fr/la-prospection-commerciale-par-courrier-electronique-sms-mms-et-automate-dappel)

B2C electronic marketing generally relies on prior consent, subject to defined exceptions.

[30] FCC - AI-generated voices in robocalls · February 8, 2024 (fcc.gov/document/fcc-confirms-tcpa-applies-ai-technologies-generate-human-voices)

The FCC confirms that the TCPA applies to AI-generated voices in robocalls.

Appendix F - Milestones 2017-2026: how modern AI was built

Why this timeline: three paths converging

Models, RAG, and agentic AI have evolved in less than a decade at a pace rarely seen in the history of enterprise technology. Reconstructing that progression makes it possible to understand how the building blocks came together and why a durable architecture must remain flexible, multi-model, and capable of integrating new advances quickly. This timeline also traces the building blocks that gradually made SACHA OMON AI possible; it summarizes Salomon Bellaïche's learning and choices through that technological evolution.

01 • MODELS AND CONSUMER AI	02 • RAG AND KNOWLEDGE	03 • AGENTIC AI AND THE ENTERPRISE
What models can understand, produce, and make accessible.	What the system can retrieve, verify, cite, and update.	What the system can plan, hand off, execute, and govern.

How to read it: each milestone title links to the corresponding scientific publication, official announcement, or primary documentation. The selection focuses on versions and launches that changed public use, RAG, or enterprise production deployment; it does not attempt to list every minor update to every model.

Why the three stories do not move at the same pace

1. Models occupy public attention. Their names and versions are visible: GPT, Claude, Gemini, Grok, Llama, Mistral, or DeepSeek. Every launch becomes an event, even though an enterprise never buys only a model name.
2. RAG advances through knowledge quality. Progress centers on dense or hybrid retrieval, ranking, contextualization, graphs, evaluation, and multimodality. These advances receive less media attention but determine response reliability.
3. Agentic AI arrives later in the enterprise. Acting requires more than generating text: identity, access rights, tools, memory, human handoff, observability, evaluations, and security. These layers become truly industrialized in 2025 and 2026.

The right unit of analysis is not the number of announcements, but the moment when the model, knowledge, and capacity for action become mature enough to work together inside an organization.

PHASE I • THE FOUNDATIONS

2017-2020: Transformer, API, and external memory

This timeline begins on June 12, 2017 with the publication of "Attention Is All You Need" by a team from Google Brain and Google Research. Artificial intelligence had existed for decades, but this work marks a turning point in modern generative AI because it introduced the Transformer architecture, which later became a foundation for BERT, the GPT family, and many large language models. Between 2017 and 2020, three decisive conditions converged: an architecture capable of processing language at scale, programmatic access through APIs, and external document memory consulted at answer time.

IN PLAIN ENGLISH • WHAT IS A TRANSFORMER?

A Transformer is a neural-network architecture that uses an attention mechanism to determine which parts of a text or sequence matter most in relation to one another. Compared with the recurrent architectures that dominated at the time, it can process far more relationships in parallel during training. This capability makes scaling and contextual understanding easier and has become one of the technical foundations of modern LLMs.

Original document: "Attention Is All You Need," arXiv, June 12, 2017, Google Research · ChatGPT, OpenAI, November 30, 2022.

01 • MODELS AND THE PUBLIC

June 2017 Google Brain and Google Research: Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin introduce the Transformer, an attention-based architecture. Open dissemination of this work allows the Transformer to become the common foundation for major model families that follow.

June 2018 OpenAI: first GPT

Generative pretraining followed by adaptation to multiple tasks launches the GPT lineage.

Oct. 2018 Google Research : BERT

Bidirectional understanding from Transformers changes information retrieval and language representation.

Feb. 2019 OpenAI: GPT-2

Coherent generation without task-specific training makes the potential of general-purpose models visible.

May 2020 OpenAI: GPT-3

Scaling shows that one model can perform many tasks from instructions and a few examples.

02 • RAG AND KNOWLEDGE

Apr. 2020 Meta AI : Dense Passage Retrieval

Dense retrieval learns to match a question with relevant passages, a decisive building block for future RAG systems.

May 2020 Lewis et al.: Retrieval-Augmented Generation

RAG formalizes the combination of a model's parametric memory with external document memory searched at response time.

03 • AGENTIC AI AND THE ENTERPRISE

June 2020 OpenAI: API opens

Programmatic access brings models into products and enterprise systems: the model becomes a callable layer rather than only a research object.

WHAT THE ENTERPRISE SHOULD REMEMBER

Between 2017 and 2020, the essential building blocks fall into place: a model that understands context, external knowledge it can consult, and API access that lets it be integrated into applications. The complete agent does not yet exist, but its fundamental dependencies are already assembled.

PHASE II • INSTRUCTION, RETRIEVAL, AND ACTION

2021-2022: the model learns to search, follow intent, and act

The three threads begin to converge. Models follow human intent more effectively, retrieval becomes part of the answer, and early reasoning-action frameworks explicitly describe an agentic loop.

01 • MODELS AND THE PUBLIC

Jan. 2022 OpenAI: InstructGPT

Training with human feedback improves instruction following and prepares the conversational form that will be adopted at scale.

Nov. 30, 2022 OpenAI: public launch of ChatGPT

The conversational interface makes generative AI understandable to the general public and sharply accelerates enterprise demand.

02 • RAG AND KNOWLEDGE

Dec. 2021 OpenAI: WebGPT

The model learns to use a browser, consult sources, and produce answers with references: a bridge between retrieval, action, and documented generation.

03 • AGENTIC AI AND THE ENTERPRISE

May 2022 MRKL Systems: routing between models and external modules

The model is presented as a router capable of calling calculators, databases, and specialized services instead of producing everything by itself.

Oct. 2022 Yao et al.: ReAct

The model alternates reasoning, action, and observation in a loop; this structure becomes one of the clearest foundations of modern agentic AI.

WHAT THE ENTERPRISE SHOULD REMEMBER

ChatGPT reveals the power of the conversational interface; ReAct and MRKL show what it still lacks for the enterprise: choosing an action, using a tool, observing the result, and trying again under control.

PHASE III • THE ECOSYSTEM OPENS UP**2023: multiple LLM families, tools, and the first enterprise services**

2023 was not just about GPT-4. Claude, Llama, Mistral, and Gemini established a multi-model market, while RAG and tool calling became accessible capabilities for developers and enterprises.

01 • MODELS AND THE PUBLIC**Feb. 24, 2023 Meta: LLaMA**

A family of open models accelerates research and makes it possible to deploy models in controlled environments.

14 Mar. 2023 OpenAI: GPT-4

GPT-4 brings multimodal capabilities and stronger execution of complex instructions in ChatGPT and via API.

14 Mar. 2023 Anthropic: Claude

Claude becomes available through a conversational interface and API, creating a second major proprietary model family for organizations.

July 18, 2023 Meta and Microsoft: Llama 2

The license permits research and commercial use, significantly expanding enterprise choice.

September 27, 2023 Mistral AI : Mistral 7B

A European provider releases a compact model under the Apache 2.0 license, important for cost, sovereignty, and local deployment.

Dec. 6, 2023 Google DeepMind: Gemini 1.0

Gemini is designed natively as a multimodal family spanning mobile devices through the largest workloads.

02 • RAG AND KNOWLEDGE**Sept. 13, 2023 AWS: preview of Agents and Knowledge Bases for Amazon Bedrock**

AWS combines models, enterprise sources, embeddings, a vector database, and RAG in an integrated cloud service.

17 Oct. 2023 Self-RAG

Retrieval becomes adaptive: the model can decide when to search, evaluate passages, and critique its own output.

Nov. 28, 2023 AWS: Knowledge Bases for Amazon Bedrock available

Managed RAG becomes an enterprise service connected to models and orchestration mechanisms.

03 • AGENTIC AI AND THE ENTERPRISE**Feb. 9, 2023 Toolformer**

The model learns to select and call external tools through APIs, a core capability of an agent that can act.

June 13, 2023 OpenAI: function calling

Models can generate structured calls to application functions, connecting conversation with business systems.

September 25, 2023 Microsoft Research : AutoGen

An open framework enables multiple agents, humans, and tools to participate in coordinated conversations.

28 Nov. 2023 AWS: Agents for Amazon Bedrock

The service decomposes a request into steps, calls enterprise APIs, and consults Knowledge Bases through RAG.

WHAT THE ENTERPRISE SHOULD REMEMBER

By 2023, the enterprise can choose among several model families, connect them to its documents, and call its own functions. This is the moment when the conversational prototype starts becoming a business architecture.

PHASE IV • CONVERGENCE**2024: multimodality, corrective RAG, graphs, and early standards**

In 2024, models see, hear, and reason better. At the same time, RAG becomes corrective, contextualized, and graph-based. Enterprises discover agents connected to CRM, data, and real tools.

01 • MODELS AND THE PUBLIC**4 Mar. 2024 Anthropic: Claude 3**

Haiku, Sonnet, and Opus models provide vision capabilities and different tradeoffs among speed, cost, and capability.

April 18, 2024 Meta: Llama 3

The open ecosystem advances in reasoning, code, and multilingual understanding.

May 13, 2024 OpenAI: GPT-4o

The same model processes text, vision, and audio in real time, bringing AI closer to web, phone, and voice interfaces.

June 21, 2024 Anthropic: Claude 3.5 Sonnet

The model improves reasoning, vision, and coding-task execution at an intermediate cost level.

July 23, 2024 Meta: Llama 3.1 405B

A very large open model strengthens the option to choose between proprietary services and controlled deployments.

Aug. 13, 2024 xAI: Grok-2

xAI establishes a competing family focused on conversation, code, and reasoning, later accessible by API.

September 12, 2024 OpenAI: o1-preview

A model family dedicated to deeper reasoning demonstrates the value of allocating more compute at answer time.

Dec. 11, 2024 Google: Gemini 2.0

Google presents a model designed for the agentic era, with native tool use and multimodal outputs.

02 • RAG AND KNOWLEDGE**29 Jan. 2024 CRAG : Corrective Retrieval-Augmented Generation**

An evaluator judges the quality of retrieved documents and triggers corrective strategies when retrieval is insufficient.

Apr. 2024 Microsoft Research : GraphRAG

Knowledge graphs derived from documents enable answers to global questions that often escape classic chunk-based retrieval.

September 19, 2024 Anthropic: Contextual Retrieval

Each chunk receives document-level context before indexing to improve the relevance of hybrid search.

Dec. 16, 2024 Microsoft: GraphRAG 1.0

GraphRAG reaches a structured release for developers and complex document-synthesis use cases.

03 • AGENTIC AI AND THE ENTERPRISE**Sept. 17, 2024 Salesforce: Agentforce launch**

Agents enter sales, service, marketing, and commerce, connected to CRM and Data Cloud.

25 Nov. 2024 Anthropic: Model Context Protocol (MCP)

An open protocol standardizes connections among assistants, data, and external tools.

Dec. 19, 2024 Anthropic: Building Effective Agents

The model augmented by retrieval, tools, and memory is established as the elementary building block of an agentic system.

WHAT THE ENTERPRISE SHOULD REMEMBER

In 2024, the quality of context changes: corrective retrieval, graphs, chunk contextualization, and the first connection standards. The agent becomes credible because it is better documented and better connected to systems.

PHASE V • MOVING INTO PRODUCTION**2025, January-June: reasoning, retrieving, and acting in real systems**

In the first half of 2025, reasoning models multiply. Major vendors ship SDKs, protocols, and managed RAG engines: the focus shifts from the chatbot to the complete architecture.

01 • MODELS AND THE PUBLIC**20 Jan. 2025 DeepSeek : R1**

An open reasoning model accompanied by distilled models significantly increases competition on price, deployment, and research.

Feb. 19, 2025 xAI: Grok 3

xAI evolves Grok toward advanced reasoning and prepares agent use cases equipped with tools.

Feb. 24, 2025 Anthropic: Claude 3.7 Sonnet

Hybrid reasoning makes it possible to choose between a fast answer and extended thinking within the same model.

25 Mar. 2025 Google: Gemini 2.5

Google introduces a family of reasoning models designed for complex and multimodal tasks.

April 5, 2025 Meta: Llama 4

The new open generation becomes natively multimodal and adopts a mixture-of-experts architecture.

April 16, 2025 OpenAI: o3 and o4-mini

Reasoning models use tools directly in ChatGPT, including web search, files, Python, and vision.

May 22, 2025 Anthropic: Claude 4

Opus 4 and Sonnet 4 explicitly target coding, advanced reasoning, and agents.

02 • RAG AND KNOWLEDGE**15 Jan. 2025 Google Cloud : Vertex AI RAG Engine**

The ingestion, chunking, indexing, retrieval, and generation cycle becomes a managed service for enterprise developers.

June 13, 2025 Anthropic: multi-agent research system

Dynamic multi-step research is distinguished from static RAG: the agent adapts its searches to the evidence found and synthesizes the results.

03 • AGENTIC AI AND THE ENTERPRISE**March 11, 2025 OpenAI: Responses API, tools, and Agents SDK**

Web search, file search, computer use, and tracing become unified components for building agents.

April 9, 2025 Google: Agent2Agent Protocol (A2A)

A2A proposes an open language that allows agents from different vendors to discover one another and collaborate.

May 27, 2025 Mistral: Agents API

Mistral provides an agents API connected to connectors, tools, code, and document libraries.

WHAT THE ENTERPRISE SHOULD REMEMBER

In the first half of 2025, the boundary between model, RAG, and agent shifts. Models reason and use tools; platforms expose retrieval and tracing; protocols begin making agents interoperable.

PHASE VI • AGENTIC PLATFORMS**2025, July-December: models, knowledge, and governance converge**

The second half of 2025 brings complete platforms: execution environments, memory, identity, tools, evaluations, observability, and policies. The model becomes a replaceable component within a broader system.

01 • MODELS AND THE PUBLIC**July 9, 2025 xAI: Grok 4**

Grok 4 integrates native tool use and real-time search, features directly useful to agents.

Aug. 5, 2025 OpenAI: gpt-oss

OpenAI releases two open-weight models focused on reasoning and tool use, suitable for controlled deployments.

Aug. 7, 2025 OpenAI: GPT-5

GPT-5 more fully unifies fast responses, reasoning, and tool selection in ChatGPT and the API.

18 Nov. 2025 Google: Gemini 3

The new generation emphasizes reasoning, multimodality, and execution of complex tasks.

Dec. 1, 2025 DeepSeek: V3.2

DeepSeek continues the convergence of reasoning, general generation, and capabilities useful to agents.

Dec. 2, 2025 Mistral: Mistral 3

Mistral Large 3 and several compact models under the Apache 2.0 license strengthen the choice among cloud, private, and local deployment.

02 • RAG AND KNOWLEDGE

July 30, 2025 Google: Gemini Embedding

Embeddings are presented as a foundation of RAG and context engineering for future agents.

Aug. 14, 2025 AWS: agentic RAG in Amazon Q Business

Retrieval becomes a planned journey: the agent selects among several navigation strategies instead of relying on a single static search.

Sept. 29, 2025 Anthropic: context engineering for agents

System quality depends on selecting and maintaining context, not only on the wording of a prompt.

03 • AGENTIC AI AND THE ENTERPRISE

July 16, 2025 AWS: Amazon Bedrock AgentCore preview

AWS provides modular infrastructure for running agents with memory, identity, gateway, tools, and observability.

Sept. 17, 2025 Salesforce: first Agentic Enterprise Index

Usage data published by Salesforce shows agents moving from prototypes into service conversations and real operations.

6 Oct. 2025 OpenAI: AgentKit

OpenAI brings together visual design, connectors, evaluations, and deployment of agent interfaces.

Oct. 13, 2025 AWS: AgentCore available

Agentic runtime, memory, identity, and observability become production services with private networking and declarative infrastructure.

Dec. 9, 2025 Anthropic: MCP joins the Agentic AI Foundation

MCP becomes an open-foundation project; Anthropic reports more than 10,000 active public servers and multi-provider adoption at that time.

WHAT THE ENTERPRISE SHOULD REMEMBER

The second half of 2025 marks the real shift to production: isolated execution, memory, identity, observability, evaluation, and policies join models and knowledge. Architecture now matters more than an LLM ranking.

PHASE VII • INDUSTRIALIZE UNDER CONTROL

2026, January-April: identity, policies, evaluations, and multimodal RAG

By early 2026, the challenge is no longer only to make an agent capable. Its rights must be limited, its paths observed, quality evaluated, and access extended to multimodal knowledge: text, images, audio, video, and documents.

01 • MODELS AND THE PUBLIC

Feb. 5, 2026 Anthropic: Claude Opus 4.6

The model supports longer agentic tasks and adds a one-million-token context window in beta.

Feb. 17, 2026 Anthropic: Claude Sonnet 4.6

Long reasoning, computer use, and agent planning become available in a more economical tier.

Feb. 19, 2026 Google: Gemini 3.1 Pro

Google improves complex reasoning and tool use in its Pro family.

5 Mar. 2026 OpenAI: GPT-5.4

The model targets long, structured professional work with tool use and document analysis.

April 23, 2026 OpenAI: GPT-5.5

The new generation improves knowledge tasks, programming, and extended agentic journeys.

April 24, 2026 DeepSeek : V4 Preview

DeepSeek experiments with a new generation of very-long-context models designed for reasoning and agent use cases.

02 • RAG AND KNOWLEDGE

March 10, 2026 Google: Gemini Embedding 2 preview

A single vector space connects text, images, audio, video, and documents, preparing truly multimodal RAG.

April 6, 2026 AWS: hybrid search for an agentic assistant

Semantic and lexical search are combined, then used inside an agentic loop with quality control.

Apr. 30, 2026 Google: Gemini Embedding 2 available

Agentic RAG can retrieve and align content in more than 100 languages across multiple modalities.

03 • AGENTIC AI AND THE ENTERPRISE

Jan. 13, 2026 Anthropic Labs: Cowork, Skills, and MCP

Anthropic brings together prototypes intended to make Claude a collaborator connected to workplace tools and knowledge.

Mar. 3, 2026 AWS: AgentCore Policy available

An agent's rights over its tools are evaluated by a policy layer external to the agent's code, with granular and auditable controls.

March 11, 2026 OpenAI: computer environment in the Responses API

Agents receive an isolated execution environment for manipulating files, code, and tools while preserving an operational trace.

Mar. 31, 2026 AWS: AgentCore Evaluations available

Quality can be measured continuously from production traces and during version changes.

April 15, 2026 OpenAI: another evolution of the Agents SDK

The SDK consolidates execution, tools, guardrails, context memory, and multi-agent orchestration.

April 22, 2026 Google Cloud : Gemini Enterprise Agent Platform

The platform brings together models, creation, orchestration, identity, security, observability, and governance for fleets of agents.

WHAT THE ENTERPRISE SHOULD REMEMBER

By early 2026, security and evaluation are no longer appendices. They become distinct services governing every tool call. At the same time, RAG moves beyond text alone and begins to understand the full range of enterprise media.

PHASE VIII • STATE OF THE ART AS OF AUGUST 19, 2026**2026, May-August: fast models, managed agents, and interoperable standards**

As this White Paper closes, the most important developments are no longer simple model records. They reduce the cost of agentic loops, make tools interchangeable, strengthen controls, and bring agents closer to the enterprise's real channels.

01 • MODELS AND THE PUBLIC**May 19, 2026 Google: Gemini 3.5 Flash**

A fast model designed for coding and agentic tasks becomes the default engine for consumer experiences such as AI Mode.

May 22, 2026 Mistral: Medium 3.5 and Work mode

Mistral combines a mid-tier model with remote agents, MCP connectors, and human controls for enterprise work.

June 9, 2026 Anthropic: Claude Fable 5 and Mythos 5

The Claude family reaches a new threshold in reasoning, agentic endurance, and specialized use.

June 26, 2026 OpenAI: GPT-5.6 Sol preview

OpenAI presents a new generation focused on research, coding, and complex problem solving.

June 30, 2026 Anthropic: Claude Sonnet 5

Sonnet 5 is presented as the most agentic Sonnet yet, capable of planning and using a browser and terminal.

July 9, 2026 OpenAI: GPT-5.6

The model generalizes Sol's advances and targets high-level research, programming, and knowledge tasks.

July 21, 2026 Google: Gemini 3.6 Flash

The model reduces the tokens and tool calls required for agentic loops at a lower price than the previous generation.

July 24, 2026 Anthropic: Claude Opus 5

Opus 5 improves long-form knowledge and engineering tasks and targets more stable agentic execution.

Aug. 6, 2026 OpenAI: GPT-5.6 Sol and Luna in ChatGPT

Sol advances for Plus and Pro plans; Luna becomes the default model for Free and Go plans, showing how quickly new generations spread into mass-market use.

02 • RAG AND KNOWLEDGE

July 16, 2026 AWS: Managed Knowledge Base for enterprise search

AWS clarifies the spectrum between a fully managed knowledge base and a custom-built RAG pipeline for agents.

July 28, 2026 MCP, specification 2026-07-28

The specification adopts a stateless core, cacheable results, and extensions for interfaces and long-running tasks in order to scale in the cloud.

03 • AGENTIC AI AND THE ENTERPRISE

May 19, 2026 Google: Managed Agents in the Gemini API

A single API call can launch an agent that reasons, uses tools, and executes code in an isolated Linux environment.

June 17, 2026 Google: Agentic Resource Discovery

An open specification targets discovery and verification of tools, skills, and agents across the web.

June 17, 2026 AWS: Guardrails in AgentCore Policy

Tool-access policies are supplemented by protections against prompt injection and exposure of sensitive data.

July 22, 2026 OpenAI: Presence for enterprise agents

OpenAI presents a layer intended to deploy and supervise agents connected to internal knowledge and tools.

July 28, 2026 Google: Managed Agents with budget controls and tool audits

Managed environments add control hooks, token caps, and scheduled triggers.

July 30, 2026 AWS: Bedrock Agents Classic enters maintenance mode

AWS directs new projects toward AgentCore, signaling the replacement of first-generation managed agents with a more complete production platform.

Aug. 6, 2026 Google and partners: Agent Plugins 1.0

A vendor-independent format packages skills, instructions, and MCP servers into portable modules.

Aug. 7, 2026 Salesforce: Agentic Enterprise Index 2025-2026

Salesforce reports an almost threefold increase in activated agents among the organizations studied and documents their move to scale.

WHAT THE ENTERPRISE SHOULD REMEMBER

As of Aug. 19, 2026, the market is converging toward managed, multi-model, interoperable agents. Durable value no longer comes from a single provider: it comes from proprietary knowledge, rules, channels, and the quality of the handoff between AI and human teams.

Legal notice

Publisher

SACHA OMON AI, a company organized under the laws of Israel. Publisher and Editor-in-Chief: Salomon Bellaïche. Contact: sacha@sacha-omon.com • www.sacha-omon.ai.

Copyright

This White Paper is protected by Israeli copyright law (Copyright Act, 5768-2007) and, through the Berne Convention, in every country where it is read. Protection arises automatically upon creation; no registration is required. It covers the text, titles, tables, diagrams, structure, and layout.

What you may do - and what is prohibited

PERMITTED	PROHIBITED WITHOUT WRITTEN PERMISSION
Read, print, and share the file as-is.	Copy the document, in whole or in substantial part.
Quote short excerpts while crediting the author and www.sacha-omon.ai .	Republish it on another website or medium.
Link to the document.	Modify, translate, or sell it.
Cite it as a reference in research or journalism.	Remove the author's name or source attribution.

Table 58: Terms of use. Permission requests: sacha@sacha-omon.com.

Search engines and artificial-intelligence systems may index and quote this document provided they identify the source.

Sources, trademarks, and liability

Figures and studies cited come from their original publishers, identified in the appendices, and remain their responsibility. Information is current as of August 19, 2026. Trademarks cited belong to their respective owners, and mention of them does not imply any partnership.

This document is informational. It does not constitute legal, financial, or investment advice, a contractual commitment, or a guarantee of results. Decisions made based on it remain the reader's responsibility.

© 2026 SACHA OMON AI. All rights reserved. International Edition 2026, Edition No. 1 • U.S. English version, Review No. 1 dated September 1, 2026.

Engage, convert and **transform** your inbound prospects into customers with Sacha.

Sacha connects to your business, learns your offers, rules and knowledge, and engages instantly with your prospects on **Web**, **WhatsApp** and **Phone** — 24/7, in 40+ languages.



Instant response
in under 2 seconds



Automatic
prospect qualification



**Responds on Web,
WhatsApp & Phone**



Books qualified
meetings for your team



Available 24/7
in 40+ languages

Sacha helps you capture, qualify and convert inbound demand faster. She answers questions, identifies real buying intent, handles objections and guides prospects to the next step — so your team can stay focused on closing deals.



WEB



WHATSAPP



PHONE

CONNECTS WHERE YOUR PROSPECTS ARE



Facebook



Instagram



Website



WhatsApp



Phone



**HIGHER
CONVERSION**

Convert more
inbound leads



**SMARTER
QUALIFICATION**

Identify real
buying intent



**TEAM
PRODUCTIVITY**

Focus on
closing deals



**ENTERPRISE
SECURITY**

Your data,
always protected

Book Your Personalized Demo

See how Sacha can qualify your prospects
and schedule more meetings.

Schedule a Meeting →



www.sacha-omon.ai



Web



WhatsApp



Phone