

ACTION IMMÉDIATE



LIVRE BLANC

Comment l'IA, propulsée par l'agentique et le RAG, entre maintenant dans l'entreprise

RAG — Retrieval-Augmented Generation

POURQUOI MAINTENANT ? • COMMENT ÇA MARCHE ? • MODE D'EMPLOI

Salomon Bellaïche

Fondateur de SACHA OMON AI • Rédacteur en chef du Livre blanc • Directeur de la publication

ÉDITION INTERNATIONALE 2026 • ÉDITION N° 1

VERSION FRANÇAISE • 21 AOÛT 2026

Données arrêtées au 19 août 2026 : sources intégralement référencées en annexe.

www.sacha-omon.ai



SACHA OMON AI
WE SELL FOR YOU

ÉDITION INTERNATIONALE • 2026

Livre blanc

Comment l'IA propulsée par l'agentique et le RAG entre maintenant dans l'entreprise

RAG — Retrieval-Augmented Generation

**POURQUOI MAINTENANT ?
COMMENT ÇA MARCHE ?
MODE D'EMPLOI**

Salomon Bellaïche

Fondateur de SACHA OMON AI
Rédacteur en chef du Livre blanc • Directeur de la publication

ÉDITION INTERNATIONALE 2026 • ÉDITION N° 1

VERSION FRANÇAISE • RELECTURE N° 14 • 21 AOÛT 2026

Données arrêtées au 19 août 2026 : sources intégralement référencées en annexe.

www.sacha-omon.ai

Sommaire

Ce Livre blanc explique comment l'IA entre maintenant dans l'entreprise grâce à la combinaison de l'agentique et du RAG — Retrieval-Augmented Generation — pour comprendre, accéder à la connaissance de l'entreprise et agir.

Besoin d'une définition ? Le glossaire page 114 explique simplement les termes techniques utilisés dans ce livre blanc. Les sources et articles cités sont regroupés pages 118 à 129.

OUVERTURE STRATÉGIQUE DU FONDATEUR

4

Comment l'IA, propulsée par l'agentique et le RAG, entre maintenant dans l'entreprise	4
Pourquoi ce livre blanc existe, et l'organisation augmentée	7
Pourquoi maintenant ? Hier il fallait être trouvé, aujourd'hui il faut agir	10
Le grand basculement	12
Guide de lecture : à qui s'adresse ce livre blanc	14
Les trois conditions de l'action	15

PARTIE I • POURQUOI MAINTENANT ?

26

01. Le paradoxe de l'IA d'entreprise	26
02. Les plateformes qui structurent l'IA d'entreprise	29
03. Cas clients et preuves chiffrées	38

PARTIE II • COMMENT ÇA MARCHE ?

42

A. COMPRENDRE LES BRIQUES

04. Comprendre les LLM : l'IA que vous connaissez, et ses limites	42
05. Comprendre l'IA agentique	48
06. Comprendre le RAG : Retrieval-Augmented Generation	55
07. IA agentique + RAG + LLM : pourquoi c'est la combinaison gagnante	62

B. DE LA DEMANDE À L'ACTION COMMERCIALE

08. De la demande à l'action commerciale immédiate	70
· De l'analyse à l'application : l'agent SDR IA agentique multicanal	76

A. CADRER VOTRE PROJET

Les 7 questions structurantes de votre cahier des charges	84
De vos réponses au cahier des charges avec votre propre IA	85
Le déploiement en escalier et les 8 marches	86
Les 14 décisions de votre agent IA	87
Faites rédiger votre cahier des charges par votre propre IA	88

B. DÉPLOYER DANS L'ENTREPRISE

09. Comment construire et déployer un agent IA dans l'entreprise	93
10. Gouvernance, sécurité et conformité	99
11. Les erreurs à éviter avant et pendant le déploiement	103
12. Là où la valeur se perd, et par où elle revient	104
Synthèse décisionnelle : ce que le dirigeant doit retenir et décider	105

À PROPOS DE SACHA OMON AI

108

ANNEXES DU LIVRE BLANC

113

Annexe A — Panorama des éditeurs de modèles et de leurs moteurs	113
Annexe B — Glossaire : termes et définitions	114
Annexe C — Méthodologie	117
Annexe D — Sources et articles cliquables	118
Annexe E — Sources cliquables de l'ouverture stratégique	127
Annexe F — Repères 2017-2026 : comment l'IA moderne s'est construite	130
Mentions légales	142
Après le Livre blanc	143

OUVERTURE STRATÉGIQUE DU FONDATEUR

Pourquoi ce livre existe, ce qui bascule, et les trois conditions de l'action.

Comment l'IA, propulsée par l'agentique et le RAG, entre maintenant dans l'entreprise

OUVERTURE DU FONDATEUR

SALOMON BELLAÏCHE

Fondateur · SACHA OMON AI

Rédacteur en chef · Directeur de la publication

Expert en transformation digitale et acquisition digitale

Plus de 30 ans d'expérience · Plus de 3 000 entreprises accompagnées

De la TPE aux grands comptes · Microsoft · L'Oréal



Salomon Bellaïche, fondateur de SACHA OMON AI.

CRÉDITS DU LIVRE BLANC

Réflexion, conception et ligne éditoriale : Salomon Bellaïche.

Recherche, structuration et vérification : assistées par Claude d'Anthropic et par les modèles d'OpenAI.

Le point de départ : protéger la valeur créée par l'acquisition

En 2025, Google et Meta ont déclaré ensemble 490,9 milliards de dollars de revenus publicitaires mondiaux dans leurs comptes officiels. Au premier semestre 2026, ce total atteint 273,3 milliards de dollars, contre 226,2 milliards un an plus tôt : une progression de 20,8 % et plus de 1,5 milliard de dollars par jour. Ces chiffres montrent l'ampleur des investissements consacrés à capter l'attention et à créer une intention.

Mais, pour l'annonceur, une intention n'est pas encore du chiffre d'affaires. C'est précisément après le clic, l'appel ou le message que la valeur peut se perdre. Le prospect arrive sur le site : il faut répondre immédiatement, comprendre sa demande, apporter une première réponse pertinente, qualifier son besoin et préparer le relais commercial. Si l'entreprise tarde à agir, l'intention refroidit avant même que l'équipe commerciale ne puisse intervenir.

C'est à cette étape qu'un agent commercial IA agentique intervient. Relié aux connaissances de l'entreprise et à ses équipes, il répond, dialogue, qualifie, traite les premières objections et transmet une opportunité exploitable. L'IA ne remplace pas la vente : elle protège la valeur créée par l'acquisition, tandis que l'humain conserve le jugement et la décision. Cette conviction repose sur plus de 30 ans d'expérience dans l'acquisition digitale, la génération de prospects et le pilotage du ROI : la performance se mesure jusqu'au chiffre d'affaires, pas seulement au clic.

Sources directes : Alphabet, 10-K 2025 et 10-Q S1 2026 · Meta Platforms, résultats FY 2025 et 10-Q S1 2026.

1996 : construire des systèmes d'acquisition, pas seulement des campagnes

Mon parcours entrepreneurial de référence commence en 1996, lorsque je crée ma première startup dans l'Internet, parmi les premières structures françaises consacrées à l'acquisition digitale et à la génération de demande. Depuis, j'ai conçu des dispositifs d'acquisition, de génération de prospects et d'optimisation du retour sur investissement pour plus de 3 000 entreprises, des TPE et PME aux grands comptes, dont Microsoft et L'Oréal. Cette expérience m'a appris à regarder toute la chaîne : la promesse, le média, la page d'arrivée, la demande, la réponse commerciale, la qualification et la transformation en chiffre d'affaires.

La question qui traverse ces trois décennies n'a pas changé : comment transformer un investissement marketing en demande réelle, puis cette demande en relation et en chiffre d'affaires ? Les outils publicitaires sont devenus plus puissants, plus précis et plus rapides. Pourtant, l'expérience proposée après l'acquisition reste souvent construite autour d'un formulaire et d'un rappel différé. Nous avons perfectionné la création de la demande sans toujours perfectionner la manière de l'accueillir.

Le formulaire enregistre l'intérêt ; il ne le traite pas

Un formulaire est utile pour collecter une information, mais il ne comprend ni l'intention, ni l'urgence, ni l'objection. Il ne conseille pas et ne répond pas à la question qui a déclenché la visite. L'exemple de Salesforce est spectaculaire. En août 2025, Marc Benioff, cofondateur, président et directeur général de Salesforce, a expliqué que plus de 100 millions de prospects ayant contacté l'entreprise au cours de ses vingt-six premières années n'avaient jamais été rappelés, faute de capacité humaine suffisante. Si le premier éditeur mondial de CRM rencontre cette rupture, aucune organisation ne peut considérer le sujet comme secondaire.

Source directe : Marc Benioff, The Logan Bartlett Show, épisode 149, 29 août 2025. Le cas complet est détaillé au chapitre 3.

Faire entrer l'IA dans l'entreprise

Les LLM ont ouvert une étape décisive. Grâce aux API et aux clés d'accès, leurs capacités de compréhension, de raisonnement et de formulation peuvent être reliées aux systèmes de l'entreprise. Mais l'usage individuel de ChatGPT ou de Claude, aussi impressionnant soit-il, ne constitue pas encore une intégration. Pour un dirigeant, la vraie question est différente : comment relier cette intelligence à la connaissance de l'entreprise, à ses règles, à ses canaux, à son CRM, à ses outils et à ses équipes, sans exposer les informations qui doivent rester protégées ?

C'est à partir de ce constat que j'ai créé SACHA OMON AI. Je n'ai pas choisi de commencer par un discours général sur la productivité. J'ai choisi le point où la perte est visible, chiffrable et directement reliée au chiffre d'affaires : l'acquisition digitale. L'objectif est de relier l'instant de la demande à celui d'une réponse fiable, immédiate et contextualisée, puis de transmettre à l'humain une situation déjà comprise et documentée. Pour construire cette approche, je me suis immergé pendant plus de trois ans dans différents

centres d'appels en Israël afin d'observer les conversations, les objections, les transferts et les moments précis où la connaissance manque.

Cette construction s'inscrit dans près de quatre années de travail sur SACHA OMON AI. Mon expérience de plus de trente ans dans le digital, l'acquisition, le développement et le pilotage du ROI m'a permis d'identifier le problème métier avant que toutes les briques technologiques ne soient mures, puis d'en suivre la convergence : les LLM pour comprendre et formuler, l'IA agentique pour conduire un objectif, le RAG pour ancrer la connaissance de l'entreprise, et le multicanal pour offrir une expérience continue. Les schémas et synthèses de ce livre traduisent cette réflexion et les choix de conception qui ont accompagné la mise en place de SACHA OMON AI ; ils en exposent les principes, pas les mécanismes propriétaires.

Agentique + RAG : l'agent agit avec les connaissances de l'entreprise

Il faut aussi être précis sur un mot qui prête à confusion : l'automatisation. Automatiser, c'est enchaîner des tâches selon des règles écrites à l'avance : si A se produit, faire B, puis C. Le système peut être très sophistiqué et relier plusieurs logiciels, mais le chemin reste déterminé d'avance — il exécute le processus qu'on lui a donné, il ne raisonne pas. J'ai encore récemment vu quelqu'un présenter comme « de l'IA » une simple réautomatisation de ses processus avec un outil de workflow. Ce n'est pas la même chose : automatiser, c'est exécuter.

Le chatbot ajoute une interface conversationnelle à cette même logique. Dans sa forme historique, il suit un scénario préconçu et se referme dès qu'on en sort ; avec les modèles de langage, il comprend mieux la demande et formule une réponse plus adaptée. Mais converser n'est pas agir : un chatbot peut très bien expliquer ce qu'il faudrait faire sans disposer de l'autonomie, des outils ni des autorisations nécessaires pour le faire. C'est ce type d'outil qui a laissé de mauvais souvenirs à beaucoup d'utilisateurs — au point que « chatbot » évoque, pour une bonne partie du grand public, un circuit fermé et des réponses à côté.

Un agent IA introduit une autre logique : il reçoit une demande, va chercher ce qu'il lui faut dans les connaissances de l'entreprise — son RAG — détermine une action en fonction de ce qu'il y trouve, choisit les outils autorisés, puis agit. Son chemin n'est donc pas entièrement écrit à l'avance : face à deux demandes différentes, il peut emprunter deux chemins différents parce que la situation ou les informations disponibles ne sont pas les mêmes. Un agent peut être déclenché automatiquement ; ce qui n'est plus automatisé à l'avance, c'est son chemin de décision — il y a une pleine décision, puis il agit. Cette différence — exécuter un processus, converser, ou réfléchir et décider avant d'agir — traverse tout ce livre. Elle explique aussi pourquoi l'IA qui entre aujourd'hui dans l'entreprise n'entre pas par la même porte que les automatisations ou les chatbots d'hier, et pourquoi elle ne doit pas hériter de leur réputation.

Un LLM seul maîtrise le langage, mais il ne connaît pas spontanément les offres, les tarifs, les procédures, les limites d'engagement ou l'historique d'une entreprise. Le RAG (Retrieval-Augmented Generation, ou génération augmentée par la recherche) lui apporte les contenus validés de l'organisation. La capacité agentique interprète l'objectif, prépare les étapes, choisit les outils autorisés, observe le résultat et adapte le parcours. Le LLM comprend et formule ; le RAG apporte les faits ; l'agentique conduit l'action dans le cadre défini.

Sacha Engine réunit ces trois dimensions dans un même centre nerveux. Une demande peut commencer sur le web, WhatsApp, le téléphone, l'e-mail ou, lorsque le format s'y prête, par SMS. Le contexte est conservé ; la base de connaissances est consultée ; les outils métier peuvent être mobilisés ; la réponse est diffusée sur le bon canal ; les sources, les échanges et les décisions sont tracés. Une seule connaissance d'entreprise peut ainsi soutenir l'avant-vente, le service, l'après-vente et le travail quotidien des équipes.

Une nouvelle fonction de l'entreprise, construite avec l'humain

L'IA n'a pas vocation à remplacer l'humain. Elle devient une fonction à part entière de l'entreprise : disponible, reliée à la connaissance et capable de préparer la bonne action. Elle porte la continuité, les réponses immédiates et les tâches répétitives ; elle remet aux équipes un prospect déjà compris, une demande qualifiée, une synthèse et une prochaine étape claire. L'humain conserve le jugement, l'empathie, la négociation, l'expertise et la responsabilité qui créent la confiance.

Conçu en Israël, pensé pour les marchés mondiaux

Cette vision s'inscrit naturellement en Israël, la Startup Nation. Son écosystème réunit entrepreneuriat, recherche, talents technologiques, investissement et ambition mondiale. Sa diversité linguistique constitue un terrain naturel pour une intelligence conversationnelle : de nombreuses langues coexistent dans les entreprises et les services. Israël forme également un pont entre les États-Unis, l'Europe, l'Asie, l'Afrique et les marchés de l'hémisphère Sud.

SACHA OMON AI a été conçue dans cette culture pour servir le marché israélien dans toute sa diversité et se déployer sur les marchés mondiaux avec une même exigence de qualité, de sécurité, d'interopérabilité et de robustesse. La technologie change ; le principe reste stable : comprendre chaque demande, mobiliser la connaissance juste, préparer la bonne action et transmettre à l'humain l'intégralité du contexte.

Un livre blanc pour les dirigeants et les décideurs opérationnels

Ce livre blanc n'a pas été conçu d'abord comme un manuel technique pour les DSI. Il s'adresse aux chefs d'entreprise, aux directions commerciales et marketing, aux responsables de l'expérience client, aux services et aux décideurs opérationnels qui doivent comprendre où l'IA crée réellement de la valeur. Il présente l'agentique, le RAG et les LLM dans l'ordre nécessaire pour décider : le problème métier, l'expérience attendue, l'architecture, les usages, la sécurité, la gouvernance et les résultats mesurables.

J'ai dirigé sa conception et sa ligne éditoriale à partir de mon expérience de l'acquisition et du développement de SACHA OMON AI. La recherche, la structuration et la rédaction ont été menées avec l'appui de ChatGPT d'OpenAI et de Claude d'Anthropic. Les sources sont indiquées afin que chaque lecteur puisse vérifier les faits, distinguer les annonces des résultats et construire sa propre décision. C'est aussi cela, intégrer l'IA avec exigence : lui donner une direction, une connaissance vérifiable et une responsabilité éditoriale humaine.

POURQUOI CE LIVRE BLANC EXISTE

Tout le monde utilise l'IA. Presque personne ne sait ce qu'il y a derrière.

Dans les entreprises que je rencontre, on ouvre ChatGPT tous les jours sans savoir ce qu'est réellement un modèle, comment il a été construit, qui le fabrique, ce qu'il fait de vos données, ni pourquoi il se trompe. Ce n'est pas de l'ignorance : c'est une question de rythme. Le sujet avance plus vite que la capacité de quiconque à le suivre, moi compris.

Et ce vide se remplit mal. On vend des formations, des tutoriels, des certifications qui décrivent, le jour même de leur mise en ligne, un état déjà dépassé du marché. Un dirigeant qui veut simplement comprendre se retrouve pris entre des promesses commerciales et des contenus périmés. Il finit par décider sans comprendre, ou par ne pas décider du tout.

Ce livre blanc existe pour cela : mettre entre toutes les mains — un dirigeant, un salarié, un étudiant, un agent public — ce que je n'ai pas trouvé moi-même. Une explication vérifiable de ce qu'est cette technologie, de ce qu'elle sait faire, de ce qu'elle ne sait pas faire, et de ce qu'il faut construire autour pour qu'elle produise un résultat. Chaque affirmation est datée et sourcée, afin que le lecteur puisse vérifier plutôt que croire.

MA RECOMMANDATION POUR TOUS

Ne vous formez pas d'abord : équipez-vous. Prenez des abonnements payants aux IA — vingt, trente ou cinquante dollars par mois et par personne. Les versions gratuites donnent accès à des modèles moins récents, moins capables et plus limités en usage : on y juge la technologie sur un état de l'art qui n'est plus le sien.

À ce prix, ce n'est pas une dépense : c'est l'investissement le plus rentable que vous puissiez engager sur le sujet, à titre personnel comme professionnel. Et c'est la seule façon de juger la technologie sur ce qu'elle fait aujourd'hui, pas sur ce qu'elle faisait il y a deux mois.

Méfiez-vous du reste. Les tutoriels qui vous font télécharger « leur » prompt en bas de la page ou sous la vidéo sont rarement gratuits : ils servent le plus souvent à toucher une commission d'affiliation, ou à vous vendre une formation. Et ce qu'on y enseigne a souvent deux versions de retard.

Une fois abonné, la meilleure formation reste souvent l'IA elle-même : demandez-lui directement ce qu'elle est capable de faire pour vous, à titre personnel comme professionnel. C'est plus fiable, et plus à jour, qu'un tutoriel enregistré deux versions plus tôt.

Quand vous l'utilisez, demandez-lui toujours ses sources, avec le lien qui permet de les vérifier. Une IA peut avancer une information sans aucune source ; exiger la source et le lien vous permet de contrôler ce qu'elle affirme plutôt que de la croire sur parole.

Mais qu'on ne s'y trompe pas : prendre un abonnement n'est pas construire un système. Ce n'est pas parce qu'une entreprise a acheté des licences et décidé de faire son agent elle-même qu'elle sait comment le faire fonctionner, ce qu'il faut lui donner, ni où se situe sa valeur réelle.

En 1996, tout le monde répétait qu'il fallait être sur Internet. Les entreprises ont acheté un site, et presque aucune ne savait comment cela fonctionnait, quel avantage en attendre, ni quel était le potentiel. Celles qui ont gagné ne sont pas celles qui ont acheté leur site le plus tôt : ce sont celles qui ont compris ce qu'il fallait construire autour.

Acheter une licence, ce n'est pas construire un agent. En 1996, avoir un accès à Internet, ce n'était pas construire une stratégie digitale pour son entreprise.

Trente ans plus tard, la scène se rejoue à l'identique, à une différence décisive : le rythme n'a rien de comparable. C'est le seul avantage que je revendique dans ce livre — avoir vécu la première révolution de bout en bout, du premier site à la vente en ligne, et pouvoir reconnaître dans la seconde ce qui relève de l'effet d'annonce et ce qui relève d'un changement de fond.

L'ORGANISATION AUGMENTÉE

Ce livre blanc est lui-même une preuve de ce qu'il décrit

Un mot revient partout : augmenté. Le salarié augmenté, l'entreprise augmentée. Le mot est juste, mais il est devenu si commode qu'il ne veut plus dire grand-chose. Je vais donc le définir par le seul exemple que je connaisse parfaitement : celui-ci.

Sans ChatGPT et sans Claude, ce livre blanc n'existerait pas. Les idées sont les miennes et l'expérience aussi : huit ans dans la production audiovisuelle, la presse et la musique, puis trente ans d'acquisition digitale, plus de 3 000 entreprises accompagnées, et plus de trois ans passés à appeler moi-même des prospects en centre d'appels. Mais structurer cette matière sur plus de cent pages, la vérifier source par source et la mener à ce niveau de détail, je n'y serais pas arrivé seul. J'aurais fait ce que l'on faisait avant : confier ma matière à un prête-plume et signer un texte écrit par un autre.

J'ai fait autre chose. J'ai gardé la main sur la pensée, sur la ligne éditoriale, sur chaque arbitrage — et sur la recherche : je n'ai pas confié à une IA le soin de décider quoi chercher, je lui ai indiqué où chercher. Ce que je lui ai confié, c'est la structuration et la vérification. Le résultat n'est pas un texte écrit à ma place : c'est un texte que je n'aurais pas pu produire à cette profondeur. Voilà ce que veut dire augmenté.

CE QUE MESURENT LES TRAVAUX DISPONIBLES

Au niveau de l'entreprise — près de 6 000 dirigeants interrogés entre novembre 2025 et janvier 2026 aux États-Unis, au Royaume-Uni, en Allemagne et en Australie : 69 % des entreprises utilisent l'IA, mais neuf sur dix ne constatent aucun effet mesurable sur leur productivité au cours des trois dernières années. Le gain moyen déclaré s'établit à 0,29 %. Ces mêmes dirigeants en attendent 1,4 % sur les trois années à venir.

Au niveau de la personne — sur 5 179 agents de support client, l'accès à un assistant IA conversationnel augmente de 14 % le nombre de dossiers résolus par heure, mais de 34 % pour les moins expérimentés, avec un effet minime sur les plus experts.

Au niveau de l'équipe — sur 758 consultants du Boston Consulting Group, l'IA fait accomplir 12,2 % de tâches en plus, 25,1 % plus vite, avec une qualité supérieure de plus de 40 %, et un gain de 43 % pour ceux qui se situaient sous la moyenne.

Là où un système a réellement été construit — 80 % des responsables techniques interrogés déclarent des retours économiques mesurables, et 57 % des organisations exploitent déjà des agents sur des processus à plusieurs étapes.

Sources directes : « Firm Data on AI », NBER working paper n° 34836, février 2026, révisé mars 2026 (Bloom, Davis, Barrero, Bunn, Yotzov et al.), enquête menée auprès de près de 6 000 dirigeants · Erik Brynjolfsson, Danielle Li et Lindsey Raymond, « Generative AI at Work », The Quarterly Journal of Economics, vol. 140, n° 2, 2025 · Fabrizio Dell'Acqua et al., « Navigating the Jagged Technological Frontier », Harvard Business School working paper n° 24-013, en collaboration avec le Boston Consulting Group, 2023 · « The 2026 State of AI Agents Report », enquête Material pour Anthropic auprès de plus de 500 responsables techniques, 2026 — étude commanditée par un éditeur de modèles, à lire comme telle.

Neuf entreprises sur dix ne voient aucun effet. La question est : pourquoi ? C'est ce que ce livre blanc explique.

Je n'ai pas « utilisé ChatGPT ». J'ai organisé un travail : une ligne éditoriale tenue par moi, des sources imposées, une vérification de chaque affirmation, un arbitrage humain à chaque étape. L'IA ne m'a pas donné d'idées. Elle m'a donné le moyen de sortir celles que j'avais depuis trente ans.

C'est exactement ce que ce livre demande à une entreprise de faire, à son échelle. Une organisation augmentée n'est pas une organisation qui a acheté des licences. C'est une organisation dont la connaissance est reliée, et qui répond à chaque demande au moment où elle arrive. Cela vaut aussi pour une mairie : un agent public qui délivre en trois minutes un document qui prenait trois semaines, et un citoyen qui n'a plus à revenir.

L'IA ne remplace pas le jugement. Elle en étend la portée.

POURQUOI MAINTENANT ?

L'entreprise entre dans l'ère du temps réel

Nous vivons dans une économie de l'immédiateté. Un repas, un trajet, un paiement ou un contenu peuvent être commandés en quelques secondes. Dans notre vie quotidienne, tout vient progressivement à nous au moment où le besoin apparaît.

Il y a trente ans, Internet a changé notre accès à l'information : nous allions la chercher. Nous ouvrions un moteur de recherche, parcourions des sites, lisions plusieurs pages et comparions. Avec l'IA générative, le mouvement s'inverse : nous posons une question et l'information vient à nous immédiatement. Nous attendons une réponse précise, concise, contextualisée et, lorsque le sujet l'exige, accompagnée de sources que nous pouvons vérifier.

Hier, il fallait être trouvé. Aujourd'hui, il faut agir.

Agir, ce n'est pas répondre. C'est comprendre la demande, la qualifier, apporter ce qui manque et la mener jusqu'au bout — un devis, un rendez-vous, un document, une décision — sans que la personne ait à chercher, à attendre ni à revenir. Cette idée traverse tout le livre : chaque chapitre décrit une condition nécessaire pour y parvenir.

Ce changement crée un paradoxe : dans notre vie personnelle, nous obtenons presque tout immédiatement, alors que face à une entreprise, une administration ou une collectivité, nous devons encore trop souvent parcourir des pages, chercher une rubrique, remplir un formulaire, laisser nos coordonnées puis attendre un rappel. L'information existe pourtant déjà dans l'organisation.

La nouvelle expérience consiste à faire venir la bonne information au moment exact où la question apparaît, sur le canal choisi : web, voix, WhatsApp, e-mail ou SMS. L'utilisateur doit pouvoir poursuivre l'échange, changer de canal, demander l'intervention d'un humain ou décider quand et comment il souhaite être recontacté. L'immédiateté doit aller avec la fiabilité, la continuité du contexte et le choix.

Cette attente concerne toute personne qui entre en relation avec une organisation : le visiteur qui découvre une offre, le prospect qui compare, le client qui cherche une réponse, l'utilisateur qui a besoin d'un service ou le citoyen qui veut comprendre une démarche. Tous ne souhaitent pas prendre rendez-vous. Souvent, ils veulent d'abord une réponse fiable, une orientation claire ou la confirmation qu'ils sont au bon endroit.

REPÈRE 2026

88 % des clients attendent des réponses plus rapides qu'un an auparavant ; 74 % attendent un service disponible 24 heures sur 24 et 7 jours sur 7.

Source : Zendesk, CX Trends 2026.

Hier, Internet nous a appris à chercher. Aujourd'hui, l'IA nous habitue à recevoir immédiatement. L'agentique relie la réponse à l'action.

Quatre publics, une même attente : maintenant

PUBLIC	BESOIN AU MOMENT OÙ IL APPARAÎT	EXPÉRIENCE ATTENDUE
Visiteur	Comprendre rapidement une offre ou un service	Une réponse claire, immédiate et adaptée à sa question
Prospect	Comparer, préciser son besoin, lever une objection	Une conversation utile sans attendre le rappel d'un commercial
Client	Obtenir une information, un suivi ou une solution	La continuité du contexte et le bon relais humain si nécessaire
Usager ou citoyen	Accéder à un droit, une procédure ou un service	Une orientation disponible, compréhensible et traçable

Tableau 1 : L'immédiateté ne signifie pas répondre à tout sans discernement ; elle signifie être présent au moment exact où la demande existe.

La nouvelle expérience proposée par SACHA OMON AI est celle d'une disponibilité continue. Sur le web, WhatsApp, le téléphone ou l'e-mail — et, lorsque le message court s'y prête, par SMS — une personne peut poser sa vraie question, recevoir une information issue des connaissances validées de l'organisation, poursuivre l'échange dans son contexte et choisir ensuite la bonne étape. Le service est accessible 24 heures sur 24 et 7 jours sur 7 ; il n'enferme pas chaque demande dans un rendez-vous et n'oblige pas l'utilisateur à recommencer son histoire à chaque canal.

L'IA agentique, le RAG (Retrieval-Augmented Generation, ou génération augmentée par la recherche) et les LLM rendent cette expérience possible ensemble. L'agent conduit le parcours, le RAG apporte les faits de l'organisation et le modèle comprend puis formule. L'humain intervient là où son jugement, son expertise, son empathie ou sa responsabilité créent le plus de valeur. La technologie étend ainsi la qualité et la disponibilité du service ; elle ne remplace pas la relation.

LA NOUVELLE PROMESSE DE SERVICE

Être présent au moment exact où le besoin apparaît. Apporter immédiatement une information précise, concise, contextualisée et vérifiable. Conserver le contexte d'un canal à l'autre. Laisser à l'utilisateur le choix de poursuivre, d'agir, de parler à un humain ou de décider quand et comment il souhaite être recontacté.

LE GRAND BASCULEMENT

L'IA est entrée dans les usages. Elle doit maintenant entrer dans l'entreprise.

L'IA générative s'est d'abord déployée auprès du grand public. Elle est devenue une interface quotidienne pour écrire, rechercher, traduire, apprendre, comparer et créer. Cette diffusion est spectaculaire : OpenAI indiquait plus de 900 millions d'utilisateurs hebdomadaires de ChatGPT en 2026. L'entreprise bénéficie de cette familiarité, mais elle n'a pas encore achevé le passage entre l'usage individuel et l'intégration opérationnelle.

900 M+

Utilisateurs hebdomadaires de ChatGPT

OpenAI, 2026

88 %

Des organisations utilisent l'IA dans au moins une fonction

McKinsey, State of AI, novembre 2025

7 %

Ont pleinement déployé l'IA à l'échelle de l'entreprise, toutes fonctions confondues

McKinsey, State of AI, novembre 2025

94 %

N'ont pas encore obtenu de déploiement global ayant un impact significatif sur le résultat opérationnel

McKinsey, 2026

Ces quatre chiffres ne mesurent pas la même chose : l'usage déclaré, le déploiement à l'échelle, l'industrialisation d'un système agentique et l'impact sur le compte de résultat sont quatre seuils distincts. Le chapitre 1 les détaille un à un.

Sources directes : OpenAI, « Scaling AI for everyone » · McKinsey, « AI at work but not at scale » · McKinsey, « Capturing Central Europe's AI opportunity ».

Ces chiffres ne se contredisent pas. Ils distinguent l'usage de l'intégration. Une organisation peut autoriser des collaborateurs à utiliser un assistant, expérimenter quelques cas et déclarer qu'elle utilise l'IA, sans avoir encore relié cette intelligence à ses connaissances, à ses droits d'accès, à ses systèmes métiers, à ses canaux et à ses indicateurs économiques. Le grand changement des prochaines années sera ce passage de l'outil personnel au système d'entreprise.

2026-2035 · LE MARCHÉ QUI S'OUVRE

Les entreprises ne veulent plus seulement des modèles puissants. Elles veulent des systèmes qui produisent des résultats.

Le marché investit déjà pour préparer cette bascule. Gartner prévoit 2 590 milliards de dollars de dépenses mondiales liées à l'IA en 2026, en hausse de 47 %. Ce montant couvre l'ensemble des dépenses liées à l'IA, matériel et services compris ; la dépense strictement agentique, capacités intégrées comprises, est estimée par le même cabinet à 201,9 milliards de dollars pour 2026 (chapitre 1). Le cabinet précise par ailleurs que la dépense reste dominée par les fournisseurs et les hyperscalers : le potentiel d'investissement des entreprises n'est pas encore pleinement engagé. Parallèlement, la part des applications d'entreprise intégrant des agents spécialisés pourrait passer de moins de 5 % en 2025 à 40 % fin 2026.

95 % DES PILOTES ÉTUDIÉS SANS IMPACT MESURABLE

Le rapport publié par le Project NANDA du MIT s'appuie sur plus de 300 initiatives publiques, 52 entretiens organisationnels et 153 réponses de dirigeants. Il estime qu'après 30 à 40 milliards de dollars investis, 95 % des pilotes intégrés étudiés n'avaient produit aucun impact mesurable sur le compte de résultat. Les auteurs identifient moins un déficit de puissance des modèles qu'un déficit de mémoire, de contexte, d'apprentissage et d'intégration aux processus.

Source primaire et statut : conclusions préliminaires, Project NANDA, juillet 2025, « The GenAI Divide: State of AI in Business 2025 ».

Ce résultat ne signifie pas que l'IA générative ne crée aucune valeur. Il montre que l'outil individuel ne devient pas spontanément un système d'entreprise. Le RAG apporte le contexte et la connaissance vérifiable ; l'agentique relie cette intelligence aux objectifs et aux outils ; l'humain conserve l'analyse critique, le jugement, la créativité et la décision. Ensemble, ces dimensions augmentent la capacité de réflexion, de production, d'analyse et d'action de l'organisation. Elles constituent une réponse architecturale aux causes observées, et non une garantie automatique de résultat.

USAGE ACTUEL	ÉTAPE D'ENTREPRISE	VALEUR RECHERCHÉE
Assistant individuel	Agent relié aux règles et aux outils	Un service disponible et mesurable
Documents dispersés	Base de connaissances gouvernée par RAG	Des réponses fiables et vérifiables
Canaux séparés	Mémoire et contexte multicanaux	Une expérience sans rupture

Sources directes : Gartner, « 2,59 T\$ de dépenses IA en 2026 » · Gartner, « 40 % des applications avec agents spécialisés ».

Dans le scénario haut de Gartner, l'agentique pourrait représenter plus de 450 milliards de dollars de revenus de logiciels d'entreprise en 2035. Cette projection ne garantit pas le succès de chaque projet. Elle indique le sens du marché : la valeur se déplacera des modèles utilisés isolément vers les architectures capables de comprendre une demande, rechercher la connaissance juste, agir sous contrôle et travailler avec l'humain.

100 millions de prospects n'avaient jamais été rappelés chez Salesforce

En août 2025, Marc Benioff, cofondateur, président et directeur général de Salesforce, a révélé que plus de 100 millions de prospects ayant contacté son entreprise au cours des 26 années précédentes n'avaient jamais été rappelés, faute de capacité humaine suffisante. Même le plus grand éditeur de CRM au monde n'échappait pas au problème que ce document documente. Quatre mois plus tard, Salesforce annonçait l'acquisition de Qualified, éditeur d'agents SDR IA. L'opération a été finalisée le 1^{er} avril 2026. Dans son Form 10-Q pour le trimestre clos le 30 avril 2026, Salesforce indique que la juste valeur de la contrepartie transférée pour Qualified était de 1,2 milliard de dollars, principalement 1,1 milliard de dollars en numéraire (chapitre 8, paragraphe 8.2).

Source directe : Marc Benioff, The Logan Bartlett Show, épisode 149, 29 août 2025.

GUIDE DE LECTURE

À qui s'adresse ce livre blanc et comment l'utiliser

Il s'adresse d'abord aux dirigeants, aux directions commerciales et marketing, aux directions des systèmes d'information et aux responsables de la transformation qui évaluent le passage à l'échelle de l'IA générative. Il s'adresse aussi aux directions des ressources humaines, aux responsables du service client et aux décideurs du secteur public, dont les enjeux sont les mêmes avec d'autres mots.

Mais il n'est réservé à personne. Étudiants, enseignants, chercheurs, journalistes, consultants, collaborateurs de toute fonction, entrepreneurs qui démarrent, ou simplement curieux qui veulent comprendre ce que l'IA change réellement dans une organisation : ce document a été écrit pour être compris sans prérequis technique. Chaque terme est défini dans le glossaire, chaque affirmation est datée et sourcée, et chaque chiffre renvoie à sa publication d'origine. Vous n'avez besoin de rien d'autre que de curiosité.

Ce livre s'adresse aussi à celui ou celle qui n'a pas encore le titre pour porter le sujet, mais qui veut être la personne qui fait entrer l'IA dans son organisation : un salarié qui pressent que son entreprise doit s'y mettre et cherche comment, un étudiant qui veut arriver en entretien avec une compréhension réelle plutôt qu'un vernis de vocabulaire. Et cette transformation ne s'arrête pas à la porte de l'entreprise. Elle entre aussi dans les mairies, les régions et les administrations, où elle change la relation entre le citoyen et son institution de la même façon qu'elle change celle entre un client et une entreprise. Du dirigeant au citoyen, en passant par le décideur et le salarié : ce livre a été écrit pour être utile à chacun, au niveau où il se trouve.

Le document reste volontairement agnostique sur le plan technologique : il décrit des motifs d'architecture, des données publiques et des retours d'expérience, avant d'illustrer leur application au cas d'usage commercial au chapitre 8.

Comment utiliser ce livre blanc

Le document peut se lire de façon continue, mais il peut aussi être utilisé comme outil de décision selon le rôle du lecteur.

BESOIN	PARCOURS RECOMMANDÉ
Découvrir le sujet sans prérequis	Ouverture du fondateur, repères 2017-2026, puis le glossaire en annexe B
Comprendre les concepts	Repères 2017-2026, puis chapitres 4 à 7
Évaluer le marché	Chapitres 1, 6 et 7, avec le focus Israël et international
Développer l'action commerciale et le service	Chapitre 8 : avant-vente, après-vente et multicanal
Préparer un déploiement	Chapitres 9 à 11, puis la grille des 14 décisions, pages 87 et 88
Arbitrer et décider	Conclusion, puis synthèse décisionnelle finale

Tableau 2 : Six portes d'entrée, selon ce que vous cherchez. Aucune ne suppose de compétence technique préalable.

TROIS RUPTURES CHANGENT LES RÈGLES

Ma lecture du nouveau rapport entre les personnes et les entreprises

Depuis 1996, je travaille sur la même question : comment une entreprise rencontre-t-elle une demande au bon moment ? Internet a d'abord donné aux personnes un pouvoir nouveau : chercher elles-mêmes. Trente ans plus tard, l'IA ouvre une autre étape. Nous ne voulons plus seulement chercher. Nous formulons un besoin et nous attendons que l'information, le service ou l'action viennent jusqu'à nous.

Ce basculement paraît discret parce qu'il s'installe dans la vie quotidienne. Pourtant, il change profondément notre seuil de tolérance. Commander un repas, réserver un trajet, effectuer un paiement, traduire un texte, comparer plusieurs options ou résumer un dossier peuvent désormais se faire en quelques gestes, parfois en une phrase.

Et c'est là que je vois un paradoxe. Nous pouvons obtenir en quelques secondes une synthèse complexe, mais nous devons encore souvent remplir un formulaire pour savoir si une voiture est disponible, si une assurance correspond à notre situation, si un logement répond à nos critères ou simplement pour obtenir un renseignement auprès d'une organisation. La technologie a changé plus vite que l'organisation du service.

MA CONVICTION

Le vrai basculement n'est pas seulement : « l'IA répond mieux ». Il est : « je formule mon besoin, et j'attends que le système organise immédiatement la bonne réponse, le bon service ou la bonne action ».

Ces trois ruptures définissent les nouvelles conditions de l'action : agir au bon moment, par le bon intermédiaire, et avec le droit de le faire.

01

Action immédiate — l'exigence du temps réel

Les attentes se forment et se décident dans l'instant. L'entreprise qui répond en différé perd la valeur de l'intention.

02

Action déléguée — les agents personnels arrivent

Chacun dispose bientôt d'un agent qui comprend, anticipe et agit en son nom. L'interface se déplace du site ou de l'application vers l'agent personnel.

03

Action autorisée — la loi et le consentement changent les règles

Transparence, consentement et contrôle deviennent la norme. La capacité d'agir avec l'autorisation du client devient un avantage compétitif.

Ces trois mouvements racontent la même histoire : la valeur se déplace vers le moment où l'intention apparaît. Les entreprises qui sauront être utiles, légitimes et immédiates à cet instant gagneront la relation.

On ne perd plus les clients faute de visibilité. On les perd faute d'action.

C'est cette grille de lecture que je propose avant les chapitres techniques. L'agentique, le RAG, les LLM, les canaux et la gouvernance ne sont pas une collection de technologies : ils sont les briques d'une nouvelle disponibilité de l'entreprise.

SUR QUOI S'APPUIE CETTE ANALYSE ?

L'immédiateté devient une norme — Zendesk, CX Trends 2026.

Les agents personnels arrivent — Google I/O 2026, « Welcome to the agentic Gemini era » ; Meta, « The Future is for Everyone » ; Microsoft, « Human-led, agent-operated ».

Le consentement et la preuve deviennent structurants — Légifrance, article L223-1 du Code de la consommation ; FCC, « AI-generated voices in robocalls ».

Sources et liens complets : annexe, pages 127 à 129.

La fin des cycles de vente différés. L'avènement de la vente en temps réel.

Internet a d'abord révolutionné l'accès. Une personne pouvait trouver elle-même une information qui supposait auparavant un déplacement, un appel, un catalogue, un intermédiaire ou du temps. Nous avons appris à taper une requête, ouvrir des liens, parcourir des pages et comparer.

Le smartphone et l'économie à la demande ont ensuite révolutionné la vitesse. Le geste est devenu plus court : je touche un écran et un trajet arrive, un repas est préparé, un paiement part, un colis est suivi. La valeur vient aussi de la disparition de l'attente inutile et de la visibilité donnée entre la demande et le résultat.

L'IA générative ajoute un troisième mouvement. Au lieu de chercher plusieurs pages puis de construire moi-même une réponse, je formule l'objectif en langage naturel. Le système cherche, rapproche, résume et explique. Les agents ajoutent une étape supplémentaire : ils peuvent continuer après la réponse, utiliser des outils et préparer une action. Google décrit Gemini Spark comme un agent personnel disponible 24 heures sur 24 ; OpenAI fait évoluer ChatGPT vers des tâches longues et multi-étapes ; Microsoft parle d'organisations « human-led, agent-operated ».

Zendesk mesure en 2026 un signal cohérent avec ce mouvement : 88 % des répondants attendent des réponses plus rapides qu'un an auparavant et 74 % associent l'IA à une disponibilité permanente. Ce n'est pas une loi universelle du comportement, mais un indicateur clair de ce que l'IA est en train de rendre normal.

HIER	AUJOURD'HUI	DEMAIN
Je cherche	Je demande	Je délègue une partie de la recherche ou de l'action
Je parcours des pages	Je reçois une synthèse	Un agent surveille, compare, alerte ou prépare la suite
Le site est l'interface	La conversation devient l'interface	L'agent peut devenir l'intermédiaire

Tableau 3 : Trois âges de l'accès à l'information et au service.

Internet nous a donné l'accès. Le mobile nous a donné l'instant. L'IA nous donne la conversation. L'agentique commence à nous donner la délégation.

Sources de cette page : Zendesk, CX Trends 2026 · Google I/O 2026 et Google, « The Gemini app becomes more agentic » · Microsoft, « Human-led, agent-operated » · OpenAI, ChatGPT Work et Workspace agents. Références complètes et liens : annexe, pages 127 à 129.

RUPTURE 01 · LE FORMULAIRE CHANGE DE NATURE

Le web n'a pas inventé le formulaire. Il l'a numérisé.

Bien avant Internet, les entreprises utilisaient déjà des coupons-réponses, des fiches clients et des formulaires papier. Le web a simplement transposé cette logique à l'écran : nom, prénom, téléphone, e-mail, âge ou autres informations utiles, puis « envoyer ».

Cette mécanique a rendu la collecte plus rapide, mais elle n'a pas fondamentalement changé la relation. La personne donne ses informations ; l'entreprise les reçoit ; puis la personne attend qu'un commercial, un conseiller ou un service la recontacte.

L'IA permet aujourd'hui de franchir une étape très simple : si une personne est déjà prête à nous donner des informations sur elle, pourquoi ne pas lui demander immédiatement ce dont elle a besoin ?

De quoi avez-vous besoin ?

La personne peut alors expliquer sa demande avec ses propres mots. L'agent IA comprend le besoin, pose seulement les questions nécessaires pour le préciser, utilise les informations déjà fournies et peut apporter immédiatement une première réponse issue des connaissances de l'entreprise.

Le changement est important : l'entreprise ne collecte plus d'abord des coordonnées pour comprendre le besoin plus tard. Elle commence par comprendre le besoin au moment même où il est exprimé. Cela permet de mieux accrocher l'intention, de répondre à la vraie question de la personne et de construire progressivement les informations nécessaires à la suite de la relation.

MA CONVICTION

Le formulaire ne disparaît pas. Il ne doit simplement plus obliger la personne à attendre pour expliquer ce qu'elle veut.

Nom, e-mail, téléphone : très bien. Mais la question la plus importante devient : « De quoi avez-vous besoin ? »

Et demain, cette demande ne sera peut-être même plus formulée directement par la personne : son propre agent IA pourra commencer à la préparer pour elle.

RUPTURE 02 · LES AGENTS PERSONNELS ARRIVENT

Les agents IA personnels connaissent notre contexte et commencent à agir pour nous

Le deuxième basculement est plus profond encore : l'utilisateur ne sera bientôt plus toujours seul face au site ou au service. Son agent pourra préparer une partie de la recherche et de l'action. Google emploie désormais explicitement l'expression « personal AI agent » pour Gemini Spark, conçu pour fonctionner en permanence en arrière-plan et agir sous la direction de l'utilisateur.

Meta suit une logique différente mais possède un avantage considérable : la conversation existe déjà dans WhatsApp, Messenger et Instagram. En juin 2026, Meta indiquait plus d'un million d'entreprises utilisant déjà Business Agent et plus d'un milliard de fils actifs avec des entreprises chaque jour sur ces trois services. WhatsApp dépasse par ailleurs 3 milliards d'utilisateurs dans plus de 180 pays.

C'est la raison pour laquelle je regarde Meta avec attention. Une publicité Facebook ou Instagram peut déjà amener une personne directement dans une conversation. Si l'IA personnelle et l'IA d'entreprise progressent dans les mêmes interfaces, la frontière entre publicité, demande, service et action devient beaucoup plus courte.

PROJECTION ÉDITORIALE

Demain, une entreprise ne recevra plus seulement des personnes. Elle recevra aussi des demandes préparées ou transmises par les agents qui travaillent pour elles.

Cette phrase est une projection, pas un fait accompli. Mais les briques sont désormais visibles : contexte personnel, mémoire, recherche, outils, voix, messagerie, permissions et agents d'entreprise. Ce qui compte n'est pas de prédire quel acteur gagnera. Ce qui compte est de comprendre que l'interface de la demande elle-même est en train de changer.

Sources de cette page : Google I/O 2026 ; Google, « The Gemini app becomes more agentic » ; Meta, Business Agent ; Meta, « WhatsApp message translations ». Références complètes et liens : annexe, pages 127 à 129.

RUPTURE 02 · SUITE — UN MOUVEMENT BEAUCOUP PLUS LARGE

Google, Meta, OpenAI, Microsoft, Grok, Apple, Amazon : chacun ouvre une porte vers le même futur

Il serait réducteur de parler seulement de trois entreprises. Les grands écosystèmes avancent par des chemins différents mais convergent vers la même idée : l'IA devient plus utile lorsqu'elle connaît un contexte, utilise des outils et peut poursuivre un objectif.

Microsoft pousse cette logique dans le travail et décrit un modèle où l'organisation reste dirigée par l'humain mais opérée en partie par des agents. OpenAI fait évoluer ChatGPT vers des workflows longs, des agents de workspace et des systèmes capables d'utiliser le web et des outils.

xAI et Grok investissent fortement dans l'agentique : outils de recherche, usage d'outils, agents vocaux, mémoire persistante dans des agents personnels comme Hermes et construction de parcours en temps réel. Apple relie Siri AI au contexte personnel à travers les messages, les e-mails et les photos, et étend les actions dans les applications.

Amazon possède un autre avantage : le commerce, la logistique, Alexa et AWS. Andy Jassy dit s'attendre à des milliards d'agents dans toutes les entreprises et tous les métiers. Ce niveau d'engagement compte parce qu'il montre que l'agent n'est plus seulement pensé comme une interface de conversation mais comme une nouvelle unité de travail logiciel.

CE QUI CONVERGE

Les interfaces diffèrent : moteur de recherche, messagerie, suite bureautique, téléphone, système d'exploitation, commerce, assistant vocal. La direction est la même : comprendre le contexte, utiliser des outils, poursuivre un objectif.

Sources de cette page : Microsoft, « Human-led, agent-operated » et Microsoft 365 Copilot ; OpenAI, ChatGPT Work, Workspace agents et Operator ; xAI, outils agentiques et voix ; Apple, Siri AI ; Amazon, Andy Jassy on generative AI. Références complètes et liens : annexe, pages 127 à 129.

CEUX QUI BÂTISSENT LE MARCHÉ

Ils parlent de travail, d'action, d'infrastructure et de contrôle. Pas seulement de chatbots.

Pour comprendre une rupture, je préfère regarder ce que construisent les entreprises mais aussi écouter ceux qui engagent leurs capitaux, leurs infrastructures et leur réputation dans cette transformation. Le vocabulaire change, mais une même tension revient : donner davantage de capacité d'action à l'IA tout en conservant une responsabilité humaine claire.

Jensen Huang, fondateur et directeur général de NVIDIA, ne parle plus seulement de modèles : en mai 2026, il affirme que l'IA agentique est arrivée, qu'elle réalise du travail productif, génère de la valeur et se déploie rapidement dans les entreprises et les industries. NVIDIA vend évidemment l'infrastructure qui rend cette croissance possible ; son discours doit donc être lu avec ce contexte. Mais le signal industriel est net.

Mark Zuckerberg pousse le raisonnement vers l'individu, avec une vision de superintelligence personnelle disponible pour chacun. Meta développe en parallèle Business Agent côté entreprise. Cette symétrie est importante : un agent côté personne, un agent côté organisation et, entre les deux, des surfaces de conversation déjà utilisées à très grande échelle.

Jeff Bezos revient lui aussi directement dans l'IA avec Prometheus, cofondée avec Vik Bajaj. En juin 2026, la société a annoncé une levée de 12 milliards de dollars pour une valorisation d'environ 41 milliards. Son objectif n'est pas de créer un chatbot supplémentaire mais de développer des outils d'IA pour l'ingénierie, la conception et le monde physique. TechCrunch et le Financial Times convergent sur ces chiffres et ce positionnement.

Les livres récents donnent encore une autre profondeur. Reid Hoffman parle de « superagency », l'augmentation de notre capacité d'action par l'IA. Mustafa Suleyman place le contrôle au centre de The Coming Wave. Eric Schmidt, avec Henry Kissinger et Craig Mundie, interroge dans Genesis une IA qui s'interpose davantage entre l'humain et la réalité.

LE FIL ROUGE

Le débat sérieux ne porte plus seulement sur la qualité d'une réponse. Il porte sur la distribution de l'action : qu'est-ce que l'agent peut faire, sur quelle connaissance, avec quelles limites, et à quel moment l'humain décide ?

Sources de cette page : NVIDIA, résultats Q1 FY2027 ; Meta, « The Future is for Everyone » et Business Agent ; Prometheus / Jeff Bezos, TechCrunch et Financial Times ; Reid Hoffman, Superagency ; Mustafa Suleyman, The Coming Wave ; Kissinger, Schmidt et Mundie, Genesis. Références complètes et liens : annexe, pages 127 à 129.

LE SIGNAL MATÉRIEL · L'AGENT SORT DE LA FENÊTRE DE CHAT

Certains agents IA disposent déjà de leur propre ordinateur

Un autre signal, encore émergent, mérite d'être regardé : certains acteurs ne vendent plus seulement un logiciel d'agent. Ils commencent à vendre une machine conçue pour qu'un agent reste actif en permanence.

À New York, The Tokenry commercialise sous la marque Otto un petit ordinateur appelé Otto One. Le produit est présenté comme une machine « always-on » pour agents IA, préchargée avec OpenClaw. Le modèle 8 Go / 256 Go est annoncé à partir de 599 dollars. La société indique une expédition prévue en janvier 2027.

Il faut être précis sur ce que cela prouve et ce que cela ne prouve pas. Otto One n'exécute pas encore les grands modèles entièrement en local : les identifiants, fichiers et tâches sont gérés sur la machine, tandis que l'inférence principale reste dans le cloud. Ce n'est donc pas la preuve que l'agent général autonome tient aujourd'hui dans un boîtier. C'est plutôt le signe qu'un nouveau marché commence à considérer l'agent comme un travailleur logiciel permanent ayant besoin de sa propre machine, de sa propre identité et de son propre environnement d'exécution.

CE QUE JE RETIENS

Le signal Otto est intéressant non parce qu'il annonce la fin du cloud, mais parce qu'il matérialise une idée : l'agent n'est plus seulement une conversation ouverte dans un onglet. Il devient un processus qui reste là, observe, exécute et attend la prochaine instruction.

Ce type de produit devra encore prouver sa sécurité, sa fiabilité, son modèle économique et son utilité réelle. Mais son existence même montre que l'agentique commence à produire une catégorie matérielle identifiable. À surveiller, donc, comme signal de marché, pas comme standard déjà établi.

Sources de cette page : Otto, Hardware ; Otto, Presale ; Otto, Contact (The Tokenry). Références complètes et liens : annexe, pages 127 à 129.

PROJECTION · L'AGENT PERSONNEL FACE À L'ENTREPRISE

Demain, votre client pourra envoyer son agent IA parler à votre entreprise

Imaginons une demande déjà banale : « Je cherche une assurance habitation pour cet appartement, avec telle couverture et tel budget. » Aujourd'hui, la personne ouvre plusieurs sites, remplit des formulaires, attend des devis et compare elle-même. Demain, une partie de ce travail pourra être déléguée à son agent personnel.

L'agent peut connaître les critères que la personne a accepté de lui confier, surveiller le marché, éliminer les offres incompatibles et demander les précisions manquantes. Le même scénario vaut pour un véhicule, un voyage, un artisan, un logement ou un service local.

La question change alors de côté : que trouve l'agent en face ? S'il ne trouve qu'une page marketing et un formulaire, il lui manque les faits opérationnels : prix à jour, disponibilité, règles, documents nécessaires, délais, créneaux, limites d'engagement et conditions dans lesquelles un humain doit intervenir.

Je ne prétends pas que tous les agents vont demain négocier automatiquement entre eux. Ce serait dépasser les preuves disponibles. Mais les briques nécessaires apparaissent déjà : agents personnels persistants, outils, voix, mandats, permissions, mémoire et agents d'entreprise.

PARCOURS D'AUJOURD'HUI	PARCOURS AGENTIQUE PLAUSIBLE
La personne cherche sur plusieurs sites	La personne exprime ses critères une fois
Elle reconstruit elle-même les informations	Son agent rapproche les informations accessibles
Elle remplit plusieurs formulaires	L'agent demande seulement ce qui manque, dans le cadre autorisé
Elle attend un rappel	Le système d'entreprise peut répondre immédiatement ou organiser le bon relais humain

Tableau 4 : Ce que change l'arrivée d'un agent du côté du client.

Un agent personnel puissant ne sert à rien si l'entreprise en face ne sait pas exposer une connaissance fiable, répondre et agir proprement.

RUPTURE 03 · LA LOI ET LE CONSENTEMENT CHANGENT LES RÈGLES

Quand le client demande à être contacté, l'entreprise doit pouvoir le prouver

La troisième rupture est réglementaire. Les règles diffèrent selon les pays, les canaux et les secteurs ; il faut donc éviter les raccourcis. Mais une direction est claire : une entreprise doit savoir pourquoi elle contacte une personne, sur quelle base, pour quel objet et avec quelle preuve.

France, 11 août 2026. Le Code de la consommation bascule le démarchage téléphonique B2C vers le consentement préalable. Le professionnel doit pouvoir démontrer ce consentement.

Pour certains secteurs, notamment la rénovation énergétique et l'adaptation du logement, le décret du 23 juillet 2026 donne un exemple très concret de la valeur d'une demande entrante explicite : lorsqu'un consommateur demande à être rappelé, le professionnel doit pouvoir justifier cette demande, respecter le périmètre demandé et intervenir dans le délai prévu, notamment cinq jours ouvrables dans le cas visé.

Il faut aussi distinguer les canaux. Le 11 août 2026 ne crée pas une interdiction nouvelle et générale de tous les e-mails et SMS : pour la prospection électronique B2C, le consentement préalable était déjà le principe, avec des exceptions encadrées.

Aux États-Unis, la FCC a confirmé que les voix générées par IA entrent dans le champ des règles du TCPA applicables aux voix artificielles ou préenregistrées. Là encore, le détail dépend du type d'appel et de son objectif.

CE QUE CELA CHANGE

Le point stratégique n'est pas : « toute prospection va disparaître ». Le point est : la preuve de l'intention et du droit de poursuivre la relation prend de la valeur.

Dans ce contexte, une demande entrante possède une provenance, une date, un canal, un objet, parfois un consentement et parfois une limite de temps. Une architecture agentique sérieuse doit pouvoir conserver ces éléments et les faire respecter pendant l'échange.

Sources juridiques : Légifrance, article L223-1 du Code de la consommation ; Légifrance, décret n° 2026-662 du 23 juillet 2026 ; CNIL, prospection commerciale par courrier électronique, SMS, MMS et automate d'appel ; FCC, « AI-generated voices in robocalls ».
Références complètes : annexe, pages 127 à 129.

LES CANAUX · LÀ OÙ LA DEMANDE EXISTE

Web, WhatsApp, voix : trois portes d'entrée vers le même besoin

Le canal n'est plus un simple support. C'est l'endroit où l'intention apparaît. Une demande peut venir d'une recherche Google, d'un clic sur une publicité Meta, d'un message WhatsApp ou d'un appel. Trois portes d'entrée sont particulièrement fortes pour la conversation immédiate : le web, WhatsApp et la voix. L'e-mail et le SMS jouent davantage un rôle de continuité, de document et de confirmation.

Le web reste central, mais l'IA permet de renverser une logique ancienne : au lieu d'obliger l'utilisateur à traduire sa question en navigation, il peut poser la question dans ses mots et le système va chercher l'information adaptée dans la connaissance de l'entreprise.

WhatsApp est différent parce que la conversation est déjà l'interface. Meta indique plus de 3 milliards d'utilisateurs dans plus de 180 pays, et plus d'un milliard de fils actifs avec des entreprises chaque jour sur WhatsApp, Messenger et Instagram. Une publicité peut donc conduire directement à une conversation, sans imposer une page intermédiaire.

La voix conserve une force particulière : lorsqu'une personne appelle, elle exprime généralement une intention immédiate. Mais la voix coûte cher à opérer humainement parce qu'elle mobilise du temps synchrone, des horaires, des langues et de la disponibilité. xAI, OpenAI et d'autres investissent désormais dans des agents capables de tenir une conversation vocale, d'utiliser des outils et de transmettre à un humain lorsque nécessaire.

CANAL	RÔLE PRINCIPAL	CE QU'IL FAUT CONSERVER
Web	Comprendre l'intention pendant la visite	Page, source, campagne, question, contexte utile
WhatsApp et messagerie	Entrer directement en conversation	Fil, origine, informations fournies, cadre de consentement
Téléphone et voix	Traiter une demande synchrone	Origine de l'appel, objet, droit de rappeler, trace
E-mail et SMS	Documenter, notifier, confirmer	Objet, historique, base d'envoi et retrait si prospection

Tableau 5 : Chaque canal a un rôle propre, mais une seule connaissance derrière lui.

Le multicanal ne consiste pas à être partout. Il consiste à comprendre la même demande, quel que soit l'endroit où elle entre dans l'entreprise.

Sources de cette page : Meta, Business Agent et « WhatsApp message translations » ; xAI, Voice API et Grok Voice Agent Builder ; OpenAI, Operator. Références complètes et liens : annexe, pages 127 à 129.

CE QUE CES TROIS RUPTURES CHANGENT POUR L'ENTREPRISE

Être visible ne suffit plus. Il faut répondre quand la demande arrive.

Pendant trente ans, le marketing digital a surtout perfectionné l'amont : être visible, être trouvé, acheter du trafic, mesurer le clic et obtenir une conversion. Ces disciplines restent essentielles. Mais elles ne suffisent plus lorsque l'utilisateur attend une réponse immédiate et lorsqu'un agent peut commencer à travailler pour lui.

Je propose donc d'ajouter une question à toutes les questions classiques d'acquisition : que se passe-t-il dans les secondes qui suivent l'apparition du besoin ? Si la réponse est « un formulaire et un rappel plus tard », l'entreprise conserve une organisation héritée du web d'hier. Si la réponse est « une conversation

fiable, contextualisée, traçable et capable de préparer l'action », elle commence à adopter le nouveau prisme.

Ce nouveau prisme ne demande pas une autonomie totale. Au contraire, les grands acteurs ajoutent tous permissions, contrôle, confirmation et supervision humaine. Google place Spark sous la direction de l'utilisateur. Microsoft parle d'organisations dirigées par l'humain. OpenAI construit des agents dans des environnements contrôlés. xAI met en avant garde-fous et outils. Meta laisse à l'entreprise le choix du moment où l'humain reprend la main.

C'est exactement le modèle que je défends : l'humain garde le dernier mot, mais il n'a plus besoin d'être le premier maillon de chaque interaction. L'IA peut porter l'immédiateté, la recherche dans la connaissance, la qualification et la préparation de l'action. L'humain intervient là où son jugement, son empathie, son expertise, sa négociation ou sa responsabilité créent la différence.

01	Capter Source, canal, date, objet, preuve de demande ou consentement si nécessaire.
02	Comprendre Intention, contexte, urgence, objection, information manquante.
03	Savoir Connaissance d'entreprise validée et gouvernée par RAG.
04	Agir Outils autorisés, étapes contrôlées, limites explicites.
05	Continuer Contexte conservé entre web, WhatsApp, voix, e-mail et SMS.
06	Transmettre Humain informé, contexte complet, décision au bon moment.

Hier, Internet nous a appris à chercher. Aujourd'hui, l'IA nous habitue à demander. Demain, des agents agiront pour nous. L'entreprise doit être prête au moment où la demande existe.

Sources de cette page : Google I/O 2026 ; Meta Business Agent ; Microsoft, « Human-led, agent-operated » ; OpenAI, ChatGPT Work ; xAI, outils agentiques. Références complètes et liens : annexe, pages 127 à 129.

AVANT D'ENTRER DANS LES CHAPITRES

Les quatre idées à garder en tête

Cette ouverture ne remplace pas les chapitres qui suivent. Elle leur donne une grille de lecture. Pour éviter de répéter les mêmes développements, les pages suivantes approfondissent chaque thème là où il devient opérationnel.

FIL ROUGE	OÙ IL RÉAPPARAÎT DANS LE LIVRE	CE QU'IL FAUT Y LIRE
Immédiateté	Pages 7-8 ; chapitre 8.1 ; chapitre 8.5	Le besoin apparaît maintenant ; la vitesse devient une composante de l'expérience et de la valeur commerciale.
Agents personnels	Chapitre 6 ; architecture du chapitre 9	L'entreprise doit rester interopérable, exposer une connaissance fiable et conserver des règles d'action.
Consentement et demande entrante	Chapitre 8.6 ; chapitre 10	La provenance, la preuve, le canal et le droit de poursuivre la relation deviennent des données opérationnelles.
Humain au centre	Chapitres 3, 5, 9 et 10	L'autonomie est bornée ; l'humain arbitre, négocie, valide et garde la responsabilité.

Tableau 6 : Quatre fils rouges et l'endroit où chacun devient opérationnel.

RÈGLE DE LECTURE

Le principe éditorial est simple : une idée forte dans l'ouverture ; une preuve, une architecture ou un cas d'usage là où le chapitre en a besoin. Pas de répétition intégrale.

Les noms des sources sont indiqués directement sur les pages concernées. Les références complètes, leurs dates et leurs liens de vérification sont regroupés dans l'annexe, pages 127 à 129. L'annexe sert à vérifier ; la page reste compréhensible sans quitter la lecture.

Les chapitres techniques restent donc inchangés dans leur rôle : ils expliquent les LLM, le RAG, l'agentique, les canaux, la gouvernance, la sécurité et le modèle de complémentarité humaine. L'ouverture explique pourquoi ces briques deviennent maintenant stratégiques.

RÈGLE TRANSVERSALE — L'ENTREPRISE DOIT ÉVOLUER AU MÊME RYTHME QUE L'IA

Les chapitres qui suivent doivent être lus avec une cinquième idée en tête : les outils, les usages et les méthodes de développement évoluent en parallèle. La bonne stratégie est donc d'avoir une vision globale de l'architecture, mais de construire par étapes, sur des besoins réels, en intégrant à chaque marche les progrès apparus entre-temps.

PARTIE I

POURQUOI MAINTENANT ?

Le paradoxe de l'IA d'entreprise, les plateformes qui basculent, et les preuves chiffrées.

01

Le paradoxe de l'IA d'entreprise

L'usage progresse partout. La performance financière, elle, reste concentrée.

Trois ans après la diffusion massive des modèles de langage, le bilan présente une asymétrie remarquable : presque toutes les organisations déclarent utiliser l'IA, très peu en tirent un effet mesurable sur leur compte de résultat. Ce chapitre expose cet écart, ses causes documentées et ce qu'il implique pour une direction générale.

CE QUI CONVERGE AU 19 AOÛT 2026

Le LLM devient une couche sélectionnable ; le RAG devient multimodal, évalué et gouverné ; l'agentique devient une couche d'orchestration dotée d'identité, de droits, de mémoire, d'outils et de traces ; les standards permettent enfin de connecter ces briques sans enfermer l'entreprise dans un seul fournisseur.

1.1 Une adoption devenue universelle

L'enquête mondiale de McKinsey publiée fin 2025 établit que 88 % des organisations interrogées utilisent régulièrement l'IA dans au moins une fonction métier, contre 78 % l'année précédente, et que 72 % recourent à l'IA générative, contre 33 % en 2024. L'intérêt pour les agents est élevé : 62 % des répondants déclarent au minimum expérimenter des agents IA.

Source : McKinsey & Company, « The state of AI in 2025: Agents, innovation, and transformation », novembre 2025.

1.2 Une transformation encore rare

Le même travail montre que près de deux tiers des organisations restent en phase d'expérimentation ou de pilote, environ un tiers ayant commencé à industrialiser. Sur le plan financier, 39 % des répondants attribuent un impact quelconque à l'IA sur le résultat opérationnel, et une fraction estimée à environ 6 % constitue le groupe des « hauts performeurs » qui rattachent plus de 5 % de leur EBIT à l'IA.

Le rapport de terrain préliminaire « The GenAI Divide: State of AI in Business 2025 », publié par le Project NANDA du MIT, examine plus de 300 initiatives publiques et s'appuie sur 52 entretiens organisationnels ainsi que sur 153 réponses de dirigeants. Après 30 à 40 milliards de dollars d'investissements, ses auteurs estiment que 95 % des pilotes intégrés observés n'avaient produit aucun impact mesurable sur le compte de résultat. Leur diagnostic porte moins sur la puissance des modèles que sur le manque de contexte, de mémoire, d'apprentissage et d'intégration aux processus. Ce signal ne constitue pas à lui seul un consensus scientifique ; il documente avec force l'écart entre une démonstration convaincante et un système d'entreprise capable de produire de la valeur.

Source primaire, conclusions préliminaires : Project NANDA, « The GenAI Divide: State of AI in Business 2025 », juillet 2025.

LA FORMULATION DU PROBLÈME

Les organisations qui obtiennent des résultats partagent trois caractéristiques : elles ancrent le modèle dans leurs propres données, elles redessinent le processus autour de l'agent plutôt que d'ajouter l'agent au processus existant, et elles mesurent la qualité en continu.

Le reste du document décrit précisément comment ces trois conditions se traduisent en architecture.

1.3 Le marché anticipe déjà cette bascule

Les prévisions divergent sur le montant et convergent sur la direction. Gartner situe la dépense agentique mondiale à 201,9 milliards de dollars en 2026 lorsque l'on comptabilise les capacités agentiques intégrées aux logiciels d'entreprise, et anticipe que 40 % des applications d'entreprise embarqueront un agent spécialisé d'ici la fin 2026, contre moins de 5 % un an plus tôt. Les cabinets qui mesurent le marché agentique « en propre » — Fortune Business Insights, Precedence Research, MarketsandMarkets, Deloitte — l'estiment entre 7 et 12 milliards de dollars pour 2025-2026, avec des taux de croissance annuels moyens compris entre 38 % et 46 %.

Le même cabinet publie l'avertissement symétrique : plus de 40 % des projets agentiques risquent l'annulation d'ici fin 2027, principalement pour valeur mal définie, coûts non maîtrisés et gouvernance insuffisante. Les deux prévisions se lisent ensemble : le marché croît vite, et la sélection y sera sévère.

CABINET	PÉRIMÈTRE MESURÉ	ESTIMATION	HORIZON
Gartner	Dépense agentique, capacités intégrées comprises	201,9 Md\$	2026
Fortune Business Insights	Marché agentic AI en propre	7,29 Md\$ → 139,19 Md\$ (TCAC 40,5 %)	2025 → 2034
Precedence Research	Marché agentic AI en propre	7,55 Md\$ → 199,05 Md\$ (TCAC 43,8 %)	2025 → 2034
MarketsandMarkets	Marché agentic AI en propre	7,06 Md\$ → 93,20 Md\$ (TCAC 44,6 %)	2025 → 2032
Deloitte	Marché agentic AI en propre	8,5 Md\$ → 35-45 Md\$	2026 → 2030
Next Move Strategy Consulting	Marché RAG	3,33 Md\$ → 81,51 Md\$ (TCAC 42,7 %)	2026 → 2035
MarketsandMarkets	Marché RAG	1,94 Md\$ → 9,86 Md\$ (TCAC 38,4 %)	2025 → 2030
The Business Research Company	Logiciels AI SDR	5,81 Md\$ → 17,58 Md\$ (TCAC 32,3 %)	2026 → 2030

Tableau 7 : Estimations publiées entre 2025 et 2026. Les écarts proviennent des périmètres retenus, non d'un désaccord sur la trajectoire.

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

La chronologie détaillée de cette construction — dix années de recherche, de produits et de standards, de l'article fondateur de 2017 jusqu'aux agents managés de 2026 — est reportée en annexe F, afin que cette partie reste consacrée à la seule question qui la fonde : pourquoi maintenant.

02

Les plateformes qui structurent l'IA d'entreprise

Quand six concurrents livrent la même architecture en dix-huit mois, ce n'est plus une hypothèse de fournisseur : c'est un standard de fait.

Entre 2025 et 2026, les principaux acteurs qui structurent l'IA d'entreprise ont tous livré la même combinaison : un moteur de raisonnement agentique, une mémoire persistante, une couche de connaissance indexée, un catalogue d'outils standardisé et un dispositif de gouvernance et d'évaluation. Chacun avec sa stratégie commerciale, tous avec la même architecture. Ce chapitre compare leurs briques et tire les conséquences pratiques de cette convergence pour un acheteur.

2.1 Amazon et AWS : industrialiser l'agent

Amazon a fait du passage en production le cœur de son argumentaire. Amazon Bedrock AgentCore est présenté comme la plateforme destinée à faire franchir aux agents l'étape entre la démonstration et l'exploitation. Les indicateurs communiqués en 2026 documentent la traction : le nombre de tâches exécutées par des agents sur AgentCore a été multiplié par quinze en six mois, et le SDK avait dépassé les deux millions de téléchargements cinq mois après sa mise en avant.

Les annonces du Sommet AWS de New York, en juin 2026, décrivent précisément l'architecture présentée au chapitre 7 :

- Fully Managed Knowledge Bases : un pipeline RAG d'entreprise managé, avec connecteurs natifs, préparation managée des formats hétérogènes et un « Agentic Retriever » dédié aux requêtes complexes en plusieurs étapes.
- AWS Context : un graphe de connaissance destiné à indiquer aux agents où trouver l'information nécessaire pour répondre juste ou décider de l'action suivante.
- Web Search sur AgentCore : un outil managé permettant d'ancrer les réponses dans une information web actuelle et citée, sans sortie de données hors de l'environnement AWS du client.
- Harness, Policy, Evaluations et Payments : respectivement l'assemblage accéléré d'un agent en deux appels d'API, des contrôles déterministes sur ce que l'agent peut faire, l'évaluation continue de la qualité sur le trafic de production, et l'exécution autonome de transactions.

Andy Jassy a résumé la logique dans sa communication de résultats du deuxième trimestre 2026 : l'assemblage fiable de ces briques constitue la difficulté réelle, et c'est précisément ce qui a bloqué de nombreux passages en production ; AgentCore existe pour fournir ces briques sous forme d'infrastructure managée. Il ajoute une prévision structurante : si les entreprises construiront leurs propres agents sur mesure, la plupart utiliseront également des services agentiques prêts à l'emploi.

Les clients cités par AWS incluent le PGA Tour, dont le système multi-agents de génération de contenu produit la couverture de tournoi dix fois plus vite, ainsi que Nasdaq, Visa et Experian pour le déploiement d'agents à l'échelle de l'entreprise.

Sources : Amazon, « AWS Summit New York 2026 », juin 2026 ; AWS News Blog ; Amazon, communication des résultats du deuxième trimestre 2026 d'Andy Jassy.

2.2 Jeff Bezos : le pari de l'IA industrielle

Le retour de Jeff Bezos à une fonction opérationnelle constitue un signal de marché à part entière. En novembre 2025, il cofonde Prometheus avec Vik Bajaj, ancien de Google X et cofondateur de Verily, et en devient co-directeur général : sa première fonction exécutive depuis son départ de la direction d'Amazon en 2021. La société démarre avec 6,2 milliards de dollars de financement.

En juin 2026, Prometheus annonce une levée de série B de 12 milliards de dollars pour une valorisation d'environ 41 milliards, auprès notamment de JPMorgan Chase, Goldman Sachs et BlackRock. L'objectif affiché : bâtir un « ingénieur général artificiel », un système capable d'accompagner la conception et la fabrication d'objets physiques complexes, du moteur d'avion au composé pharmaceutique. Bezos décrit publiquement l'ambition comme une version très moderne de la conception assistée par ordinateur, et précise que le projet est étranger à la robotique. La société opère depuis San Francisco, Londres et Zurich.

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

LA LECTURE STRATÉGIQUE

Le pari de Bezos porte sur l'IA appliquée au monde physique. Sa signification pour le sujet de ce livre blanc tient au déplacement qu'il illustre : la valeur migre de la conversation vers l'exécution. Un système qui conçoit un moteur et un agent qui conduit un cycle commercial complet relèvent du même mouvement de fond : l'IA qui produit un résultat, plutôt que l'IA qui produit une réponse.

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

ACTEUR	OFFRE AGENTIQUE	POSITION ET DIFFÉRENCIATION
Microsoft	Copilot Studio, Azure AI Foundry, Agent 365	Héritage d'identité : les agents reprennent les politiques Entra ID, Purview et Defender déjà en place. Facturation par utilisateur, prévisible. Avantage décisif dans les organisations standardisées Microsoft 365.
Google	Gemini Enterprise, Vertex AI Agent Builder et Agent Engine, Antigravity	Choix de modèles très large via Model Garden, sessions et Memory Bank pour la mémoire persistante au niveau de l'infrastructure, gouvernance et piste d'audit unifiées.
OpenAI	AgentKit, Agents SDK, Workspace Agents	Approche outil de développement : construire, évaluer et déployer vite. Écosystème de connecteurs étendu, mémoire persistante par espace de travail.
Anthropic	Claude, Model Context Protocol	Le protocole d'interopérabilité devenu standard de fait, transféré en décembre 2025 à l'Agentic AI Foundation sous l'égide de la Linux Foundation.
xAI	Grok API, outils agentiques, appel de fonctions	Famille de modèles de raisonnement accessible par API, avec appels d'outils, sorties structurées et intégration à une architecture multi-modèles.
Salesforce	Agentforce 360, Atlas Reasoning Engine, Data Cloud	Agents ancrés dans la donnée CRM. Métrique publiée en unités de travail agentique plutôt qu'en licences : un basculement de discours du siège vendu vers le travail effectué.
AWS	Bedrock AgentCore, Knowledge Bases, AWS Context	Infrastructure agentique managée, agnostique aux frameworks, avec évaluation et politiques intégrées.
Éditeurs spécialisés	Sierra, Decagon, Qualified, Intercom Fin	Agents verticaux en contact client, souvent facturés à la résolution : le modèle économique s'aligne sur le résultat.
SACHA OMON AI	Sacha Engine (moteur agentique) et RAG dédié par client	Agent SDR IA agentique multicanal unique sur web, WhatsApp, téléphone, e-mail et SMS, à contexte centralisé ; infrastructure mono-locataire ; comptes et données au nom du client.

Tableau 8 : Panorama des offres agentiques structurant le marché américain, SACHA OMON AI inclus à titre de repère de positionnement, état à mi-2026.

2.4 Trois choses que les grandes plateformes ont désormais en commun

1. La connaissance d'entreprise devient un produit d'infrastructure. Knowledge Bases managées chez AWS, Data Cloud chez Salesforce, Memory Bank et Model Garden chez Google : l'ancrage documentaire a quitté le statut de brique applicative pour rejoindre celui de service de plateforme.
2. La mémoire persistante s'impose comme critère de production. Les trois grands fournisseurs de modèles ont livré en 2026 un mécanisme de mémoire longue durée. Ce point recoupe exactement la cause d'échec identifiée par le MIT.
3. La gouvernance conditionne le déploiement. Politiques déterministes, évaluations continues sur trafic réel, journalisation et identité des agents figurent désormais au catalogue de chaque plateforme. Le marché a intégré que l'autonomie sans traçabilité reste invendable en entreprise.

CE QUE CETTE CONVERGENCE SIGNIFIE POUR UN ACHETEUR

Lorsque six acteurs en concurrence frontale livrent la même architecture en dix-huit mois, le motif cesse d'être une hypothèse de fournisseur pour devenir un standard de fait. Trois conséquences pratiques en découlent.

Le choix se déplace du modèle vers l'exploitation. Les moteurs de raisonnement se substituent les uns aux autres ; la qualité de la base de connaissance, la finesse de l'orchestration et la rigueur de l'exploitation créent l'écart durable.

La portabilité devient un critère contractuel. L'adoption d'un protocole ouvert et gouverné collectivement se vérifie, se demande et s'écrit dans un contrat.

La preuve se fait sur données réelles. Les grandes plateformes ont livré des dispositifs d'évaluation continue précisément parce que la démonstration ne prédit plus la performance en production.

2.5 La prochaine étape : les agents IA commencent à dialoguer entre eux

Les protocoles décrits dans ce chapitre ne se limitent pas à connecter un agent d'entreprise à ses propres outils. Une seconde couche de standards, déployée par les mêmes acteurs entre 2025 et 2026, vise à connecter les agents entre eux, et dessine une évolution du commerce que les directions marketing et commerciales ont intérêt à anticiper.

Google a publié le protocole Agent2Agent (A2A) le 9 avril 2025, avant d'en transférer la gouvernance à la Linux Foundation le 23 juin 2025. Là où le MCP relie un agent à ses outils, l'A2A relie les agents entre eux : découverte mutuelle, échange sécurisé d'informations, coordination d'actions. Lors du passage en version 1.0, en avril 2026, le protocole comptait plus de 150 organisations partenaires, dont AWS, Cisco, IBM, Microsoft, Salesforce, SAP et ServiceNow.

Le 16 septembre 2025, Google a complété ce socle par l'Agent Payments Protocol (AP2), une extension d'A2A conçue avec plus de soixante partenaires du paiement — Mastercard, American Express, PayPal, Adyen, Coinbase — et du commerce, dont Salesforce et Etsy. AP2 représente chaque achat exécuté par un agent sous la forme de trois « mandats » signés cryptographiquement : l'intention exprimée par l'utilisateur, le panier constitué par l'agent, et le paiement autorisé. L'objectif est de prouver, de manière vérifiable, qu'une personne réelle a bien autorisé l'action que l'agent exécute en son nom.

De son côté, OpenAI a lancé le 29 septembre 2025, avec Stripe, l'Agentic Commerce Protocol (ACP), qui alimente la fonction Instant Checkout de ChatGPT, étendue le 16 février 2026 à l'ensemble des utilisateurs américains sous le nom « Buy it in ChatGPT ». Avec plus de 900 millions d'utilisateurs actifs chaque semaine, ChatGPT devient à ce titre un canal de vente à part entière, au même rang qu'un site web.

VISION PROSPECTIVE DU FONDATEUR — HYPOTHÈSE, PAS PRÉDICTION

Cette lecture prospective est proposée par Salomon Bellaïche. Elle s'appuie sur des protocoles et des travaux réels, mais elle ne décrit ni un standard stabilisé, ni une trajectoire certaine. Le web agentique peut évoluer autrement.

SCHÉMA 1

Le web agentique

L'agent personnel découvre l'agent d'entreprise, dialogue avec lui puis peut conclure une transaction.

REGISTRE D'AGENTS OUVERT — REGISTRES PUBLICS, ANNUAIRES D'AGENTS, GPTS · DÉCOUVERTE

A	AGENT PERSONNEL — CÔTÉ UTILISATEUR Agent qui accompagne un individu : préférences, budget, historique.
B	AGENT DE L'ENTREPRISE — CÔTÉ ENTREPRISE S'appuie sur les règles métier, produits, tarifs, disponibilités.
1	DIALOGUE — PROTOCOLE A2A Questions, besoins, critères, disponibilité, informations.
2	TRANSACTION — PROTOCOLE AP2 Mandat, accord, négociation, réservation, achat.

| Réponse obtenue ou achat conclu, sans qu'un humain ne visite le site web de l'entreprise.

Figure 1 : Dans cette vision, un agent personnel découvre l'agent IA d'une entreprise via un registre ouvert, dialogue avec lui, puis peut aller jusqu'à une transaction.

Cette hypothèse converge avec plusieurs briques déjà visibles. À Google I/O 2026, Google a présenté un panier unifié qui transforme Gemini en agent d'achat personnel sur ses services, ainsi qu'un Business Agent permettant à un client de dialoguer directement avec l'agent IA d'une marque plutôt qu'avec son seul site.

En parallèle, le projet NANDA du MIT Media Lab décrit un web plus lisible par les agents IA : un annuaire ouvert, le NANDA Index, et un format d'identité vérifiable, AgentFacts, pour qu'un agent puisse authentifier un autre agent avant d'échanger avec lui.

Ma lecture de cette convergence est la suivante : demain, l'agent personnel d'un client pourrait interroger directement l'agent IA d'une entreprise pour obtenir une information, comparer une offre ou conclure un achat. C'est une direction plausible, pas une certitude. Pour l'entreprise, le point pragmatique est déjà

valable aujourd'hui : construire un agent IA exact, documenté par RAG et capable de dialoguer via des protocoles interopérables prépare plusieurs futurs possibles sans les présumer.

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

2.6 Israël : quand les grands groupes achètent, financent et construisent l'IA

Le signal à regarder n'est pas seulement le nombre de startups. Il faut regarder où les grands groupes achètent des technologies, recrutent les équipes et installent durablement leurs capacités de recherche et développement.

85 Md\$

Exportations high-tech
israéliennes en 2025
Israel Innovation Authority, 2026

≈ 15 Md\$

Capital levé en 2025
Israel Innovation Authority, 2026

84 Md\$*

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

35 %

Des investissements dirigés
vers le cœur de l'IA
Israel Innovation Authority, 2026

** Le chiffre des exits concerne l'ensemble de la high-tech israélienne, pas seulement l'IA : les plus grosses opérations de 2025 relèvent notamment de la cybersécurité. Il mesure la profondeur de l'écosystème et l'appétit des acheteurs mondiaux.*

2025-2026 : les signaux IA à retenir

OPÉRATION	CE QU'IL FAUT RETENIR
Apple et Q.AI	Acquisition confirmée en janvier 2026. Reuters estime la valeur autour de 1,6 Md\$. Q.AI travaille sur l'IA audio.
Anthropic et Decart AI	Discussions révélées le 13 août 2026 autour d'une opération pouvant atteindre environ 6 Md\$. À la date de ce Livre blanc, l'acquisition n'est pas finalisée.
NVIDIA	Run:ai, startup israélienne d'optimisation des infrastructures IA, a été acquise pour environ 700 M\$ fin 2024. En 2025, NVIDIA a lancé un projet d'expansion en Israël pouvant accueillir plusieurs milliers de salariés ; des plans ultérieurs évoquent jusqu'à 10 000.
Meta et SSI	Meta a tenté d'acquérir Safe Superintelligence, qui a une implantation à Tel-Aviv, puis a recruté son cofondateur Daniel Gross après le refus. C'est un signal de guerre des talents, pas une acquisition de SSI.
Amazon — repère structurel	Annapurna Labs n'est pas une acquisition 2025-2026 : Amazon l'a achetée en 2015. Mais cette équipe israélienne est devenue une brique centrale des puces AWS, notamment Trainium. C'est un précédent historique, pas un accord récent.

Tableau 9 : Les opérations israéliennes qui documentent l'appétit des grands acteurs pour l'IA.

Vérification : AI21 Labs n'est pas comptée comme une acquisition NVIDIA. Des discussions ont été rapportées fin 2025, puis AI21 a officiellement contesté l'existence de négociations d'acquisition en janvier 2026.

France et Italie : le profil n'est pas le même

PAYS	PROFIL OBSERVÉ
Israël	Densité de deep-tech, capital étranger et sorties vers des acheteurs mondiaux. Les acquisitions et la guerre des talents sont un signal fort.
France	109 Md€ d'engagements d'investissement IA annoncés en 2025, surtout infrastructure et centres de données. ASML a investi 1,3 Md€ dans Mistral AI ; Mistral a acquis Koyeb en 2026.
Italie	Environ 10 Md€ d'investissements en centres de données annoncés pour 2025-2026. G42 et iGenius ont lancé un grand projet de supercalcul IA ; le signal visible est davantage infrastructure et usage industriel que fusions-acquisitions de startups IA.

Tableau 10 : Trois écosystèmes, trois profils d'investissement. Comparaison indicative, non exhaustive.

Lecture : dans les sources examinées, Israël se distingue surtout par la densité de technologies revendables et l'appétit des acheteurs internationaux. La France concentre davantage de grands projets d'infrastructure et de financement ; l'Italie, de capacité de calcul et d'usages industriels.

Sources vérifiées au 19 août 2026 : Israel Innovation Authority, High-Tech Status Report 2026 ; Reuters (Apple/Q.AI, 29 janvier 2026 ; Anthropic/Decart, 13 août 2026 ; NVIDIA, 2024-2025 ; Meta/SSI, 3 juillet 2025 ; France, 9 février et 9 septembre 2025, 17 février 2026 ; Italie, 16 janvier et 16 mai 2025).

2.7 États-Unis, Israël, Europe, Asie : un même socle, quatre logiques d'adoption

SCHÉMA 2

Un marché global, quatre logiques d'adoption

Une architecture commune, adaptée aux logiques de chaque région.

ÉTATS-UNIS	ISRAËL	EUROPE	ASIE
Échelle et revenu	Startup Nation	Confiance et contrôle	Canaux et langues
Déploiement rapide, intégration CRM, mesure du pipeline et du retour commercial.	Recherche, talents, entrepreneuriat, innovation appliquée et ambition internationale dès l'origine.	Gouvernance, conformité, souveraineté des données et supervision humaine explicite.	Expériences mobile-first, messageries, voix, disponibilité continue et diversité linguistique.

Une même architecture agentique, quatre logiques d'achat.

Figure 2 : La même architecture agentique se déploie selon des logiques régionales différentes.

Aux États-Unis, l'acheteur attend une preuve de contribution au chiffre d'affaires, une intégration rapide à la pile commerciale et la capacité à monter en charge. En Israël, l'écosystème valorise la profondeur technique, l'innovation appliquée, la robustesse et l'ambition internationale. En Europe, les décisions sont fortement structurées par la gouvernance, la localisation, la propriété des données et la supervision humaine. En Asie, la diversité linguistique, le mobile, les messageries et la disponibilité sur plusieurs fuseaux deviennent centrales.

Ces attentes se complètent et forment une grille de conception globale : produire de la valeur mesurable, innover rapidement, donner confiance et s'adapter aux usages réels. La France conserve un rôle dans cet ensemble, notamment pour la conformité, le langage et certaines industries, mais elle ne doit pas définir seule la taille ni l'ambition du marché adressé.

RÉGION	QUESTION D'ACHAT DOMINANTE	RÉPONSE ARCHITECTURALE
États-Unis	Quel impact sur le pipeline, la conversion et l'échelle ?	Intégration CRM, vitesse de déploiement, mesure du chiffre d'affaires
Israël	Le système est-il robuste, différencié et exportable ?	Architecture appliquée, sécurité, interopérabilité, adaptation rapide
Europe	Qui contrôle les données, les actions et la conformité ?	Isolation, gouvernance, audit, supervision et hébergement régional
Asie	Le service fonctionne-t-il dans nos langues et nos canaux ?	Messageries, voix, disponibilité permanente, localisation linguistique et culturelle

Tableau 11 : Les logiques régionales à intégrer dans une stratégie agentique internationale.

POSITIONNEMENT SACHA OMON AI

Conçue et développée en Israël pour les marchés internationaux, SACHA OMON AI associe une logique de produit américaine — chiffre d'affaires, vitesse, intégration — à une architecture de contrôle attendue en Europe et à une expérience multicanale et multilingue adaptée aux marchés asiatiques.

2.8 Concevoir en Israël avec une architecture pensée pour l'international

Concevoir une solution destinée à plusieurs marchés implique d'intégrer dès l'origine la diversité des langues, des canaux, des règles, des systèmes métier et des exigences de gouvernance. L'architecture doit pouvoir évoluer sans reconstruire le système à chaque nouveau pays ou à chaque nouveau canal.

Israël constitue un point d'ancrage particulièrement adapté à cette logique : diversité linguistique, culture technologique, proximité avec les grands écosystèmes internationaux et conception naturellement tournée vers l'export. Cette réalité pousse à penser l'interopérabilité, la sécurité et le multilingue dès le départ.

Pour SACHA OMON AI, le marché israélien constitue le point d'ancrage. Les États-Unis, l'Europe et l'Asie représentent des marchés cibles dont les attentes sont prises en compte dans la conception : mesure de la valeur et intégration commerciale aux États-Unis, gouvernance et contrôle en Europe, diversité des langues, de la voix et des usages mobiles en Asie. Il ne s'agit pas de prétendre être déjà déployé partout, mais de construire une architecture capable de s'adapter progressivement à ces environnements.

COUCHE	SOCLE INTERNATIONAL	ADAPTATION RÉGIONALE
Connaissance	Sources validées, versionnées et gouvernées	Langue, terminologie, règles locales et droits d'accès
Intelligence	Orchestration agentique et choix multi-modèles	Modèle, latence, localisation des traitements et politique d'usage
Canaux	Web, WhatsApp, téléphone, e-mail et SMS	Canaux dominants, consentement, rythme et conventions conversationnelles
Systèmes métier	CRM, agenda, catalogue, support et journalisation	Connecteurs, hébergement, identité et processus de l'organisation
Preuve	Qualité, sources, actions, transferts et résultats	Indicateurs de chiffre d'affaires, de service, de conformité ou d'expérience

Tableau 12 : Une architecture pensée pour l'international, avec des adaptations explicites selon le marché visé et l'organisation.

L'enjeu est donc simple : concevoir dès aujourd'hui un socle qui puisse demain fonctionner dans plusieurs marchés sans perdre la connaissance, le contexte, la gouvernance ni la qualité de service. Ce qui change d'un pays à l'autre, ce sont les règles, les canaux dominants et les attentes ; le principe d'architecture reste commun.

PRINCIPE DE DÉPLOIEMENT

Un seul centre nerveux pour la connaissance, le contexte et la gouvernance ; des modèles, des connecteurs, des canaux et des règles adaptés à chaque marché.

2.9 Réutiliser les preuves d'un marché pour accélérer le suivant

Un déploiement international devient plus rapide lorsque les preuves produites dans un premier marché sont structurées pour être réutilisées. Il ne s'agit pas de copier un résultat commercial hors de son contexte, mais de conserver les actifs qui permettent de le reproduire : scénarios d'évaluation, questions de référence, sources validées, limites d'action, mesures de latence, règles de transfert humain, incidents observés et critères d'acceptation.

Cette discipline permet d'élargir les langues, les canaux, les produits et les pays en conservant un niveau de preuve constant : cas difficiles, corrections, décisions humaines et résultats deviennent une mémoire opérationnelle qui accélère les déploiements suivants.

TYPE DE PREUVE	QUESTION DÉMONTRÉE	ACTIF RÉUTILISABLE
Technique	Le système cherche-t-il, raisonne-t-il et agit-il de façon fiable ?	Jeux de tests, traces, sources, seuils de latence et taux d'erreur
Métier	Le service améliore-t-il réellement le parcours et le résultat ?	Indicateurs de qualification, conversion, résolution, coût et transfert
Gouvernance	Les données, droits et actions restent-ils contrôlés ?	Politiques, rôles, validations, journaux et procédure d'incident
Adoption	Les équipes et les clients utilisent-ils le service avec confiance ?	Retours terrain, motifs d'escalade, formation et critères de satisfaction

Tableau 13 : La preuve internationale repose sur des actifs transférables, pas sur une promesse générale.

03

Cas clients et preuves chiffrées

Les annonces ne suffisent plus. Ce chapitre ne retient que ce qui a été publié, chiffré et daté.

Les résultats publiés montrent ce qui produit déjà de la valeur ; les échecs documentés montrent ce qu'il faut corriger. Les uns et les autres sont présentés avec leur source et la réserve d'usage qui s'applique à toute communication d'éditeur. Un échec n'est pas un verdict définitif sur l'IA agentique : il doit être replacé dans la maturité des modèles, des outils et des méthodes disponibles au moment du projet.

3.1 Résultats publiés

ORGANISATION OU PÉRIMÈTRE	RÉSULTAT PUBLIÉ	SOURCE
Salesforce : base clients Agentforce	Plus de 100 M\$ d'économies annualisées cumulées et +34 % de productivité déclarés par les clients ; plus de 18 000 clients dans 124 pays.	Salesforce, page Agentforce Metrics
Salesforce : exercice FY2026	ARR Agentforce de 800 M\$, en hausse de 169 % sur un an ; 2,4 milliards d'unités de travail agentique délivrées sur Agentforce et Slack.	Résultats Salesforce FY2026, février 2026
Salesforce : site d'aide	Plus de 75 % des demandes résolues par l'agent sur un site recevant plus de 60 millions de visites annuelles.	Salesforce, étude de cas
Nexo	62 % des dossiers résolus par l'agent après une phase de reprise de dette technique et de nettoyage de données.	Salesforce, étude de cas
Safari365	Objectif d'efficacité de 15 % dépassé, résultat supérieur à 30 %, service assuré 24 heures sur 24 sur plusieurs fuseaux.	Salesforce, étude de cas
Wiley	Réduction d'environ 50 % du temps de montée en compétence des conseillers assistés par l'agent.	Benchmarks de déploiement 2026

ORGANISATION OU PÉRIMÈTRE	RÉSULTAT PUBLIÉ	SOURCE
PGA Tour (AWS)	Production de la couverture éditoriale de tournoi dix fois plus rapide via un système multi-agents.	AWS, juin 2026
Amazon Bedrock AgentCore	Tâches exécutées par les agents multipliées par quinze en six mois ; SDK téléchargé plus de 2 millions de fois en cinq mois.	AWS, 2026
Demandbase (agent SDR IA sur l'inbound)	Pipeline issu du canal doublé, rendez-vous sur comptes cibles doublés, plus de 100 heures de SDR économisées par mois, environ 80 k\$ d'économies.	Étude de cas Qualified / SaaS
Éditeur d'agents CX (Sierra)	Chiffre d'affaires annualisé estimé à 200 M\$ en mai 2026, contre environ 130 M\$ fin 2025 : croissance portée par la facturation à la résolution.	Sacra, mai 2026
Éditeur d'agents CX (Decagon)	Plus de 100 nouveaux clients grands comptes en 2025 ; chiffre d'affaires annualisé estimé à 35 M\$ en octobre 2025.	Sacra
Reporting financier exécuté par agents (multi-agents)	Délai de production ramené de 15 jours à 35 minutes ; coût par rapport réduit d'environ 2 200 \$ à 9 \$; erreurs moyennes de 3 à 0,3 par rapport.	Étude de cas éditeur / intégrateur

Tableau 14 : Résultats publiés. Les chiffres émanent des éditeurs ou de leurs clients et méritent la prudence d'usage applicable à toute communication commerciale.

3.2 Preuves académiques

- Finance. Le pipeline agentique RAG à requêtes multiples de l'IIT Madras améliore l'exactitude de 11,2 % et réduit les hallucinations de 15 % sur des jeux de questions-réponses financières standard.
- Éducation. Le système multi-agents MARAUS, évalué sur plus de 6 000 interactions réelles, atteint 92 % d'exactitude moyenne, ramène le taux d'hallucination de 15 % à 1,45 % et répond en moins de quatre secondes, pour un coût d'exploitation de deux semaines inférieur à 12 dollars.
- Science. Un cadre agentique hybride appliqué à la littérature scientifique montre que la récupération hybride constitue le socle le plus solide, le raisonnement agentique apportant l'expansion contrôlée des preuves et la synthèse ancrée.
- Secteur public. Une évaluation de fidélité au niveau de l'énoncé, conduite avec plusieurs modèles juges, met en évidence l'apport de l'attribution par outil dans les architectures multi-outils sur l'interprétabilité et la traçabilité du contenu produit.

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

3.3 Les échecs documentés : des retours d'expérience datés, pas des verdicts

Trois séries de faits publics méritent d'être exposées, parce qu'elles délimitent les conditions de réussite avec plus de précision que les succès.

CAS	FAITS PUBLICS	ENSEIGNEMENT
Klarna	L'entreprise annonce que son assistant IA réalise le travail de 700 conseillers, puis réintègre des équipes humaines après un constat public de qualité dégradée et de plaintes en hausse. Le dispositif fonctionne désormais en modèle hybride.	L'exécution agentique intégrale du contact client sans escalade humaine calibrée dégrade la marque plus vite qu'elle ne réduit les coûts.
11x	Une enquête de TechCrunch en mars 2025 relève l'utilisation de logos de non-clients et un chiffre d'affaires annualisé surévalué ; des clients rapportent une performance inférieure à celle des équipes humaines.	Un pilotage sur des données réelles vaut mieux qu'une démonstration.
Marché AI SDR	Une analyse de 100 000 e-mails appariés publiée en avril 2026 par Digital Applied mesure un taux de réponse de 4,1 % pour les e-mails générés par IA contre 5,2 % pour les e-mails rédigés par des humains, et un taux de rendez-vous pris de 0,7 % contre 1,1 %. Le taux de signalement comme spam est de 8 % pour l'IA contre 3 % pour l'humain.	La performance provient de la qualification et de l'ancrage, plutôt que du volume d'envoi.
Support client, écart annoncé et constaté	Un audit de marché publié le 15 août 2026 recense des taux annoncés par les fournisseurs de 65 % à 86 %, tandis que les données de production publiées se concentrent entre 40 % et 70 %. L'écart tient notamment aux définitions différentes de la « résolution », à la complexité des demandes et à la qualité de la base de connaissances.	La qualité de la base de connaissance détermine le résultat davantage que le choix du modèle.
Prévision Gartner	Plus de 40 % des projets agentiques risquent l'annulation d'ici fin 2027, pour valeur mal définie, coûts non maîtrisés et contrôles insuffisants.	Le cadrage de la valeur et la gouvernance constituent la première ligne du plan projet.

Tableau 15 : Cinq enseignements tirés d'échecs et d'écarts publiquement documentés.

Ce que ces échecs disent vraiment

Ces cas restent utiles, mais ils ne doivent pas être lus comme des limites immuables de l'IA agentique. Un projet lancé il y a deux ans n'a pas été construit avec les mêmes modèles, les mêmes plateformes, les mêmes capacités d'intégration ni les mêmes habitudes d'usage qu'en 2026. Trois évolutions changent la lecture des échecs.

1. Les outils évoluent. Des briques autrefois développées sur mesure — mémoire, identité, recherche, observabilité, connecteurs, contrôles — deviennent progressivement des composants de plateforme. Le travail se déplace vers l'architecture, les données, les règles et la qualité.

2. Les usages et les mentalités évoluent. Clients, collaborateurs et dirigeants s'habituent à demander une information ou une action à une IA. L'adoption n'est pas figée : elle progresse avec la familiarité, la confiance et la qualité du service.
3. La manière de développer évolue. La programmation, les tests, la documentation et certaines intégrations sont de plus en plus assistés par l'IA. Un projet trop long et trop figé peut donc consacrer du temps à construire aujourd'hui ce qu'une plateforme saura fournir plus simplement demain.

CONVICTION DU FONDATEUR — VISION GLOBALE, DÉPLOIEMENT PAR MARCHES

C'est la méthode que je propose : choisir un besoin réel, construire une première capacité, la confronter au terrain, mesurer, corriger, intégrer les progrès apparus entre-temps, puis étendre. Ce n'est pas une règle universelle ; c'est ma conviction pour éviter de figer un projet pendant que la technologie change.

PARTIE II

COMMENT ÇA MARCHE ?

A. Comprendre les briques — les LLM, l'agentique, le RAG, et pourquoi leur combinaison change tout.

04

Comprendre les LLM : l'IA que vous connaissez, et ses limites dans l'entreprise

Elle connaît le dehors de votre entreprise. Elle ne connaît pas le dedans.

Ce chapitre part d'un constat que vous pouvez vérifier vous-même en trente secondes : l'IA que vous utilisez déjà répond très bien sur ce qui est public, et se trompe dès qu'il s'agit de l'intérieur de votre organisation. Nous décrivons d'abord cette limite et ce qu'elle coûte, puis pourquoi elle existe, puis ce qu'il faut ajouter pour la lever.

4.1 Un LLM, c'est l'IA que vous connaissez déjà

Derrière le sigle LLM — Large Language Model, ou grand modèle de langage — se trouve simplement le moteur de langage qui fait fonctionner les IA que vous utilisez tous les jours : ChatGPT, Claude, Gemini, Copilot. Quand un dirigeant dit « j'utilise l'IA », il utilise un LLM.

Ce moteur fait deux choses, et il les fait remarquablement bien. Il comprend une demande formulée en langage ordinaire, sans mot-clé, sans formulaire, sans syntaxe particulière. Et il produit une réponse claire, adaptée au ton et au contexte de la question. À cela s'ajoutent la synthèse, la traduction, la rédaction, le raisonnement pas à pas et l'écriture de code. C'est une capacité réelle, et c'est ce qui explique l'adoption massive de ces outils.

Une distinction utile pour la suite du livre : le LLM est le moteur, pas le produit. ChatGPT, c'est le moteur plus une interface, une mémoire, des outils et des règles construits autour. Un agent d'entreprise suit

exactement la même logique : ce n'est pas un modèle, c'est un modèle plus ce qu'on a bâti autour. Tout ce livre parle de ce qu'il faut bâtir autour.

4.2 Le dehors et le dedans : la limite fondamentale

Un LLM a appris sur des textes publics. Il connaît donc très bien le dehors de votre entreprise : ce que dit votre site, votre plaquette, vos communiqués, les articles qui vous citent, les avis de vos clients. Interrogé là-dessus, il répond juste, et vite.

Il ne connaît pas le dedans. Il ignore qui fait quoi chez vous, comment se déroule réellement une procédure, quelles conditions vous accordez à quel client, ce qui est en stock ce matin, où en est un dossier, ce qui a été promis lors du dernier échange. Cette information n'est écrite nulle part publiquement — elle est privée, elle est vivante, et elle change tous les jours.

CE QUE L'IA SAIT DE VOTRE ENTREPRISE	CE QU'ELLE NE PEUT PAS SAVOIR
Ce que vend l'entreprise, tel que le site le présente	Les conditions réelles négociées avec chaque client
Le discours public : plaquette, communiqués, presse	Les procédures internes et les règles qui font autorité
La réputation : avis clients, forums, classements	Qui fait quoi, qui valide quoi, qui doit être prévenu
Le secteur, les concurrents, le vocabulaire du métier	Les prix appliqués, les stocks, les délais du jour
Les connaissances générales jusqu'à son entraînement	L'historique d'un client et l'état d'un dossier en cours

Tableau 16 : La frontière n'est pas technique, elle est informationnelle. Le dehors est public ; le dedans est privé, propre à l'entreprise, et il change en permanence.

Or c'est dans le dedans que se prennent les décisions commerciales. Une réponse qui ne connaît que le dehors est une réponse de brochure.

C'est la raison technique de tout ce qui suit. Une entreprise ne peut pas agir sur une demande si le système qui lui répond ignore ses règles, ses prix, ses procédures et l'état de ses dossiers. Agir suppose d'abord de savoir.

FAITES LE TEST — TRENTE SECONDES

Ouvrez ChatGPT ou Claude et posez trois questions sur votre propre entreprise.

1. Que vendez-vous ? — La réponse est bonne : l'information est sur votre site.
2. Quelle est votre procédure de remboursement au-delà de trente jours ? — Le modèle invente une réponse plausible, ou vous indique qu'il ne peut pas savoir.
3. Où en est le dossier de mon client Dupont ? — Impossible.

Vous venez de mesurer, sur votre propre cas, où s'arrête le dehors et où commence le dedans. Tout ce livre blanc traite de la deuxième partie.

4.3 Les six limites d'une IA utilisée seule dans l'entreprise

Cette frontière entre le dehors et le dedans produit six limites concrètes. Aucune ne se règle en changeant de modèle ou en prenant un modèle plus gros : elles se lèvent en ajoutant quelque chose autour du moteur. La troisième colonne indique quoi, et le chapitre du livre qui la traite.

LA LIMITE	CE QUE CELA DONNE CONCRÈTEMENT	CE QU'IL FAUT POUR LA LEVER
1. Elle ne connaît pas le dedans de votre entreprise	Réponses génériques, inutilisables face à un client ; ou invention pure lorsque la question porte sur une règle interne.	Une base de connaissance privée, gouvernée, interrogée avant chaque réponse — le RAG, chapitre 6
2. Elle ne connaît pas non plus le présent	Sa connaissance est figée au jour de son entraînement : tarifs, catalogue, effectifs, jurisprudence, stock lui échappent.	Une connaissance datée, versionnée et mise à jour — le RAG, chapitre 6
3. Elle ne se branche pas toute seule	Le modèle ne lit pas votre CRM, n'ouvre pas votre agenda, n'accède pas à vos documents tant qu'on ne l'y a pas autorisé et connecté.	Un travail d'intégration : données, droits d'accès, canaux, supervision — chapitres 9 et 10
4. Elle ne fait rien	Elle produit du texte. Elle ne réserve pas un créneau, n'écrit pas dans le CRM, ne relance pas, ne transmet pas un dossier.	Une capacité d'action encadrée — l'agentique, chapitre 5
5. Sans usage défini, elle ne produit aucun résultat mesurable	Chacun l'utilise dans son coin. La productivité reste individuelle, le processus de l'entreprise ne change pas.	Un cas d'usage cadré et mesuré — la méthode, partie III
6. Ce que vous lui donnez peut sortir de l'entreprise	Un salarié qui dépose un prévisionnel, un plan de trésorerie, un fichier client ou un contrat dans une IA grand public confie ces informations à un service que son organisation ne contrôle pas.	Un environnement dédié, où la donnée reste dans le périmètre de l'entreprise — chapitres 7 et 10

Tableau 17 : Six limites, six réponses. Aucune ne dépend de la taille du modèle.

La cinquième est la plus coûteuse, et la moins visible. Le rapport publié par le Project NANDA du MIT estime qu'après 30 à 40 milliards de dollars investis, 95 % des pilotes intégrés étudiés n'avaient produit aucun impact mesurable sur le compte de résultat. Ce n'est pas un problème de modèle : c'est un problème d'usage non défini et de système non relié.

Source primaire et statut : conclusions préliminaires, Project NANDA, juillet 2025, « The GenAI Divide: State of AI in Business 2025 ». Analyse détaillée au chapitre 1.

La sixième limite mérite un développement : c'est la plus répandue et la moins vue

Cette limite n'est pas technique, elle est organisationnelle. Partout, des salariés utilisent l'IA de leur propre initiative, avec leur compte personnel, et le plus souvent pour bien faire leur travail : préparer une note, résumer un contrat, bâtir un prévisionnel. On l'observe dans les entreprises privées comme dans le secteur public, jusque dans des administrations et des ministères. L'intention est bonne. Le problème est que les conditions d'usage diffèrent selon le compte utilisé, et que presque personne ne les lit.

CE QUE DIT LE CONTRAT QUE PERSONNE NE LIT

Sur les offres grand public, les échanges peuvent servir à améliorer les modèles. Chez OpenAI, c'est le comportement par défaut, avec une option de refus dans les réglages. Chez Google, c'est le cas tant que l'historique d'activité reste actif — le réglage initial. Chez Anthropic, depuis août 2025, l'utilisateur doit choisir explicitement à la création de son compte.

Sur les offres professionnelles — API, Team, Business, Entreprise —, ces trois éditeurs excluent contractuellement les contenus de leurs clients de l'entraînement de leurs modèles.

La différence ne tient donc ni à la technologie, ni au modèle : elle tient au contrat souscrit et au compte employé. Un même modèle, deux régimes.

Les mesures disponibles donnent l'ordre de grandeur du phénomène. Sur l'ensemble des données d'entreprise transmises à des outils d'IA, la part de données sensibles est passée de 10,7 % à 27,4 %, puis à 34,8 % en trois ans. Une analyse portant sur 22,4 millions de requêtes professionnelles y a relevé 579 113 cas contenant des données sensibles. Ce ne sont pas des fuites au sens classique : ce sont des transmissions volontaires, faites par des collaborateurs de bonne foi.

Sources directes : Cyberhaven Labs, AI Adoption and Risk Report 2025 · Harmonic Security, analyse de 22,4 millions de requêtes professionnelles, 2025 · conditions d'utilisation et politiques de confidentialité d'OpenAI, Google et Anthropic, consultées le 19 août 2026.

On ne bâtit pas un prévisionnel commercial ou un plan de trésorerie dans une IA grand public : ce que l'on y dépose, on ne maîtrise plus ce qu'il en advient.

L'inverse est vrai, et c'est l'argument le plus utile pour un dirigeant. Dans un environnement dédié à l'entreprise, la sécurité redevient maîtrisable : les données restent dans un périmètre identifié, les accès sont attribués, les usages sont tracés. À partir de là, l'outil peut être ouvert à tous les salariés sans crainte. Et chacun dispose alors de deux connaissances au lieu d'une : celle du modèle, générale, et celle de son entreprise, apportée par le RAG. C'est précisément l'effet décrit au chapitre 7.

4.4 Pourquoi ces limites existent : comment l'IA lit et comment elle répond

Pour comprendre pourquoi une IA invente plutôt que d'avouer qu'elle ne sait pas, il faut savoir ce qui se passe entre votre question et sa réponse. Deux mécanismes suffisent à tout expliquer.

Premier mécanisme : elle ne lit pas des mots, elle lit des nombres

Un réseau de neurones ne manipule ni lettres ni mots : il manipule des nombres. Avant tout traitement, un composant appelé tokeniseur découpe donc votre texte en fragments — les tokens — et attribue à chacun un identifiant numérique. Un token n'est ni une lettre ni exactement un mot : c'est un fragment fréquent de la langue. Les mots courants forment un seul token ; les mots rares, longs ou mal orthographiés sont découpés en plusieurs morceaux.

CE QUE VOUS ÉCRIVEZ	CE QUE LE MODÈLE VOIT	NOMBRE DE TOKENS
bonjour	un mot courant, gardé entier	1 token environ
anticonstitutionnellement	un mot rare, découpé en morceaux	6 à 8 tokens environ
Une page A4 de texte	environ 500 mots	650 à 800 tokens environ
Un contrat de 40 pages	le document entier envoyé au modèle	30 000 tokens environ

Tableau 18 : Ordres de grandeur indicatifs. Le découpage exact dépend du tokeniseur propre à chaque modèle et de la langue.

Conséquence rarement anticipée pour une entreprise francophone ou multilingue : à contenu identique, le français consomme généralement plus de tokens que l'anglais. Une même conversation ne coûte donc pas le même prix selon la langue.

Second mécanisme : elle prédit la suite la plus probable

Toute cette industrie repose sur une architecture décrite le 12 juin 2017 par une équipe de Google dans un article intitulé « Attention Is All You Need » : le Transformer. Son idée centrale tient en une phrase. Plutôt que de lire un texte mot après mot dans l'ordre, le modèle observe tous les éléments d'une séquence en même temps et calcule lesquels comptent le plus les uns par rapport aux autres. C'est le mécanisme d'attention, et c'est lui qui a rendu possible la qualité de langue que nous connaissons aujourd'hui.

Mais ce que le modèle produit, au fond, reste une prédiction : le fragment suivant le plus probable, compte tenu de tout ce qui précède. Il n'a pas de notion de vrai ni de faux, seulement de vraisemblable. C'est là que se joue tout le reste : lorsqu'il ne dispose d'aucune information sur votre procédure de remboursement, il ne s'arrête pas — il produit la suite la plus plausible. C'est ce qu'on appelle une hallucination, et ce n'est pas un défaut de fabrication : c'est la conséquence directe de la façon dont l'outil est construit.

Un modèle ne ment pas : il complète. Tant qu'on ne lui fournit pas le dedans de l'entreprise, il le devine.

Document original : « Attention Is All You Need », arXiv, 12 juin 2017, Google Brain et Google Research. Chronologie complète : annexe F.

4.5 Ce que vous payez : le token est aussi l'unité de vente

Cette mécanique a une conséquence économique directe. Puisque tout ce que le modèle doit savoir doit lui être transmis à chaque échange, et puisque tout est compté en tokens, le token est devenu l'unité commerciale du secteur. Les fournisseurs ne facturent ni à l'utilisateur, ni à la question, ni au mois d'usage : ils facturent au token, avec deux compteurs distincts.

1

Les tokens d'entrée

Tout ce que le système envoie au modèle : la question, mais aussi l'historique de la conversation, les documents retrouvés dans votre base de connaissance, les règles et les instructions. C'est le poste le plus lourd, et le moins visible.

2

Les tokens de sortie

Le texte produit par le modèle. Généralement facturé plus cher que l'entrée, mais en volume bien plus faible.

Deux mécanismes viennent nuancer ce calcul. La mise en cache du contexte évite de repayer plein tarif un contenu déjà transmis, ce qui compte beaucoup pour un agent qui rappelle les mêmes règles à chaque échange. Et le traitement par lots, lorsque la réponse n'est pas attendue dans l'instant, est facturé à tarif réduit. Les modalités varient d'un fournisseur à l'autre et changent souvent.

Dans un système agentique, une seule demande d'un client peut déclencher plusieurs appels au modèle : comprendre, chercher, vérifier, formuler, agir. Le coût ne dépend donc pas de la longueur de la réponse affichée à l'écran, mais du nombre d'étapes et du volume de contexte mobilisé à chacune. Le pilotage des tokens devient une décision d'architecture autant qu'une ligne budgétaire.

Le token ne représente toutefois qu'une partie de la facture. Le coût complet d'une tâche agentique — recherche, appels d'outils, voix, stockage, exploitation et supervision — est détaillé au paragraphe 8.10.

4.6 La réponse tient en deux briques

Les six limites décrites plus haut se ramènent à trois manques. Il manque au modèle la connaissance du dedans ; il lui manque la capacité d'agir ; et il manque à l'entreprise la maîtrise de ce qui sort de chez elle. Trois réponses y correspondent.

CE QUI MANQUE	L'ARCHITECTURE QUI RÉPOND	CE QU'ELLE APPORTE
La connaissance du dedans	Le RAG — Retrieval-Augmented Generation, chapitre 6	Une base privée, datée et gouvernée, interrogée avant chaque réponse : le modèle répond à partir de vos documents, pas de sa mémoire
La capacité d'agir	L'IA agentique, chapitre 5	Un objectif, un plan, des outils autorisés, une observation du résultat et une correction jusqu'au bout de la tâche
La maîtrise de ce qui sort de l'entreprise	Un environnement dédié et gouverné, chapitres 7 et 10	Des données qui restent dans un périmètre identifié, des accès attribués et des usages tracés : l'outil peut alors être ouvert à tous les salariés

Tableau 19 : Trois manques, trois réponses. Le chapitre 7 montre pourquoi leur combinaison produit autre chose que leur somme.

Le chapitre suivant présente la première de ces deux architectures, celle qui donne au modèle la capacité d'agir, avant que le chapitre 6 ne détaille la seconde, celle qui lui donne la connaissance de votre entreprise.

05

Comprendre l'IA agentique

L'IA agentique marque le passage d'une intelligence qui répond à une intelligence qui agit.

Contrairement à un modèle qui se contente de produire du texte en réponse à une instruction, un agent IA reçoit un objectif, élabore un plan, mobilise les ressources autorisées, exécute des actions et corrige sa trajectoire en fonction des résultats obtenus. Il combine raisonnement, mémoire, outils et autonomie bornée pour produire un résultat vérifiable. Ce chapitre pose les fondements de ce mode de fonctionnement.

QUATRE DIMENSIONS RECONNAISSABLES

Intentionnalité — il poursuit un objectif défini par l'utilisateur ou par le système.

Autonomie — il choisit les actions à entreprendre et se corrige lorsque c'est nécessaire.

Outillage — il utilise des outils extérieurs : recherche documentaire, API, bases de données, agenda.

Apprentissage continu — il s'appuie sur la mémoire et sur les retours d'expérience pour s'améliorer.

L'agent IA ne se contente pas de comprendre. Il agit. Il s'adapte. Il accomplit.

DÉFINITION SIMPLE

Un système agentique reçoit un objectif, comprend la situation, prépare un plan, choisit des outils, observe ce qui se passe, corrige sa trajectoire et s'arrête lorsque le résultat attendu est atteint, impossible ou doit être confié à un humain.

5.1 Ce que signifie réellement « agentique »

Dans une interaction classique, l'utilisateur donne une instruction et le logiciel renvoie une sortie. Dans une interaction agentique, l'utilisateur ou le processus exprime un état à atteindre : qualifier une demande, résoudre un incident, préparer une proposition, vérifier une éligibilité, trouver un créneau ou mettre à jour un dossier. Le système doit alors choisir plusieurs étapes et les enchaîner. La différence est moins visible dans la phrase produite que dans la responsabilité confiée au dispositif.

Le modèle de langage reste le moteur de compréhension et de raisonnement, mais il n'est qu'un composant. Autour de lui se trouvent une mémoire externe, des règles, des connecteurs, des outils, une base de connaissances, des contrôles d'autorisation, une journalisation et un mécanisme d'escalade. C'est cet ensemble qui forme l'agent. Dire qu'un LLM est un agent reviendrait à dire qu'un moteur est une voiture : il fournit l'énergie centrale, mais pas la direction, les freins, la carrosserie ni les règles de circulation.

SYSTÈME	CE QU'IL REÇOIT	CE QU'IL PRODUIT	RESPONSABILITÉ
Chatbot à règles	Une intention prévue	Une réponse ou un menu	Suivre un scénario fermé
Assistant LLM	Une instruction	Un texte, une synthèse, une proposition	Aider l'humain à décider
Workflow déterministe	Un événement déterminé	Une séquence fixe	Exécuter le workflow prévu
Système agentique	Un objectif et des contraintes	Un résultat, une action et une trace	Choisir et adapter le chemin dans un périmètre borné

Tableau 20 : Quatre familles de systèmes souvent confondues. L'agentique se distingue par l'adaptation du chemin, non par l'absence de règles.

Le terme que le marché emploie souvent pour désigner un « workflow déterministe » est « automatisation » : une architecture qui exécute un chemin fixé à l'avance, utile pour un processus stable et répétitif. Le système agentique, à l'inverse, choisit son chemin à chaque interaction et le corrige selon ce qu'il observe — c'est cette capacité de choix qui distingue les deux catégories, et non le degré de sophistication du modèle utilisé.

5.2 La boucle agentique, étape par étape

SCHÉMA 3

La boucle agentique

Un objectif métier piloté jusqu'à un résultat vérifiable.

AGENT IA CADRÉ	
1	OBJECTIF Définir l'objectif.
2	COMPRENDRE Interpréter la demande.
3	PLANIFIER Choisir les actions.
4	UTILISER LES OUTILS Exécuter avec les bons leviers.
5	OBSERVER Contrôler le résultat.
6	ADAPTER Corriger et améliorer.

| Garder le focus · Itérer · Apprendre · Supervision humaine.

Figure 3 : Une boucle agentique transforme un objectif métier en suite d'étapes observables et contrôlées.

L'objectif donne une direction. Il décrit l'état attendu et les limites à respecter. « Aider ce prospect à choisir l'offre adaptée et obtenir son accord pour un rendez-vous » est agentique ; « écrire un e-mail » est une tâche isolée.

La compréhension construit la situation. Le système rassemble le message courant, l'identité du contact, l'historique autorisé, le canal, la langue et les règles applicables. Il distingue ce qui est certain, ce qui manque et ce qui doit être vérifié.

La planification choisit un chemin. Selon la situation, l'agent peut répondre directement, demander une précision, rechercher une information, appeler un outil, proposer une action ou transmettre à un humain. Le plan n'est pas forcément long ; il doit être proportionné.

Les outils produisent un effet réel. Une recherche documentaire, un CRM, un agenda, une API de stock, une solution de paiement ou un système de ticketing deviennent des fonctions autorisées. L'agent ne simule pas l'action : il l'exécute ou prépare sa validation.

L'observation contrôle l'effet. Après chaque appel, le système vérifie le résultat : le créneau est-il encore disponible ? la source répond-elle réellement à la question ? l'enregistrement a-t-il été créé ? L'échec devient une information de pilotage.

L'adaptation évite l'entêtement. L'agent peut reformuler une recherche, choisir un autre outil, réduire son ambition, s'abstenir ou escalader. Cette capacité de correction est l'un des traits les plus importants d'un système agentique de production.

5.3 Répondre n'est pas encore agir

SCHÉMA 4

De la réponse à l'exécution

La référence n'est pas la qualité de l'IA : c'est le résultat de la tâche.

OBJECTIF MÉTIER

ASSISTANT CONVERSATIONNEL

1 • QUESTION

Poser une question.

2 • GÉNÉRATION

Échanger et comprendre.

3 • RÉPONSE

Fournir une réponse.

SYSTÈME AGENTIQUE

1 • OBJECTIF

Définir le but.

2 • PLAN + RÉFLEXION

Analyser et décider.

3 • OUTILS + CONTRÔLE

Utiliser et vérifier.

4 • ACTION + PREUVE

Agir et fournir une preuve.

Tâche cognitive → **Résultat traçable.**

Figure 4 : La frontière opérationnelle se situe entre la production d'une réponse et la conduite d'une tâche jusqu'à un résultat traçable.

Une réponse peut déjà être extrêmement utile : résumer un contrat, rédiger un message, expliquer une procédure. L'agentique commence lorsque le système porte une partie de la continuité opérationnelle. Il se souvient que la personne a déjà envoyé un document, vérifie ce qui manque, consulte une règle à jour, propose l'étape suivante, obtient une validation puis inscrit le résultat dans le système de référence.

Cette distinction évite deux erreurs. La première consiste à appeler « agent » tout assistant qui possède un bouton ou un connecteur. La seconde consiste à croire qu'un agent doit être entièrement autonome. En production, l'autonomie utile est presque toujours bornée : plus large sur les actions réversibles et documentées, plus étroite sur les engagements financiers, juridiques ou humains.

5.4 L'autonomie est un réglage, pas une idéologie

L'agentique n'impose pas de retirer l'humain de la boucle. Elle permet de choisir précisément où l'humain intervient. Une organisation peut exiger une validation systématique au lancement, passer ensuite à un contrôle par exception, puis autoriser certaines actions à faible risque lorsque les évaluations montrent une qualité stable. L'autonomie se gagne par la preuve et se retire lorsque le contexte change.

TYPE DE DÉCISION	RÔLE DE L'AGENT	RÔLE HUMAIN RECOMMANDÉ
Information factuelle sourcée	Rechercher, synthétiser, citer	Audit périodique des sources
Action réversible	Exécuter dans des limites prédéfinies	Contrôle par exception
Engagement commercial	Préparer, calculer, proposer	Valider les dérogations et les négociations
Cas sensible ou ambigu	Détecter le risque, rassembler le contexte	Décider et prendre la relation
Incident de qualité	S'abstenir, journaliser, transférer	Corriger la règle, la donnée ou le processus

Tableau 21 : La complémentarité se conçoit décision par décision.

5.5 Ce que l'agentique ne doit pas devenir

- Un synonyme d'exécution mécanique sans contrôle. Un agent de production travaille dans un périmètre explicite et sait transférer.
- Une démonstration spectaculaire sans indicateur. La réussite se mesure par la qualité du service, l'achèvement de la tâche, la satisfaction, la traçabilité et l'impact métier.
- Un modèle omniscient. L'agent ignore les connaissances propriétaires tant qu'elles ne lui sont pas fournies par une couche de connaissance gouvernée.
- Une boîte noire sans contrôle. Les appels d'outils, les sources, les actions, les autorisations et les escalades doivent être journalisés ; les raisonnements privés du modèle n'ont pas à être exposés.
- Un substitut uniforme aux équipes. Les meilleurs cas d'usage couvrent l'indisponibilité, préparent le travail, maintiennent le contexte et augmentent la capacité de service.

À RETENIR AVANT LE RAG

L’agentique apporte la capacité de poursuivre un objectif. Elle ne lui apporte pas, à elle seule, la connaissance de l’entreprise. La question suivante devient donc inévitable : sur quelles informations cet agent doit-il fonder ses décisions ? C’est précisément le rôle du RAG.

Un agent agentique est un système qui reçoit un objectif plutôt qu’une instruction, élabore un plan, sélectionne et appelle des outils, observe les résultats obtenus, corrige sa trajectoire et conduit la tâche jusqu’à un état terminal vérifiable. Le modèle de langage y tient le rôle de moteur de raisonnement ; l’agent est l’ensemble du dispositif construit autour.

5.6 Les six composants d’un agent de production

COMPOSANT	RÔLE	ILLUSTRATION COMMERCIALE
Objectif	L’état visé, exprimé en résultat métier plutôt qu’en tâche.	Qualifier la demande entrante et poser un rendez-vous qualifié.
Planification	Décomposition de l’objectif en étapes ordonnées, révisées en cours d’exécution.	Identifier le besoin, vérifier l’éligibilité, proposer un créneau, confirmer.
Outils	Fonctions appelables : recherche documentaire, API, CRM, agenda, paiement, téléphonie.	Lecture du calendrier, création d’opportunité, envoi de confirmation.
Mémoire	Mémoire courte pour la conversation, mémoire longue pour l’historique du contact.	Reprise d’un échange interrompu la veille sur un autre canal.
Réflexion	Autocritique : la réponse est-elle fondée, l’objectif est-il atteint ?	Détection d’une information manquante avant de proposer un créneau.
Garde-fous	Politiques déterministes, périmètre d’action, escalade vers l’humain.	Transfert immédiat vers un conseiller au-delà d’un seuil défini.

Tableau 22 : Anatomie d’un agent. La valeur se concentre dans l’orchestration, plus que dans le modèle sous-jacent.

5.7 Les niveaux d’autonomie en production

Cette échelle ne mesure pas l’intelligence du modèle. Elle décrit la responsabilité opérationnelle confiée au système et, en miroir, le rôle conservé par l’humain. Monter d’un niveau n’est pertinent que lorsque le périmètre, les droits, les évaluations et le transfert humain sont déjà maîtrisés.

SCHÉMA 5

L'autonomie est un réglage, pas une absence de contrôle

À chaque niveau correspondent un périmètre agent et une responsabilité humaine explicites.

NIVEAU	CE QUE LE SYSTÈME PEUT FAIRE	RÔLE HUMAIN
N0 • Assistant	Répond, synthétise et produit du contenu.	L'humain décide et exécute chaque action.
N1 • Copilote	Propose une action prête à être validée.	L'humain valide chaque action avant exécution.
N2 • Agent supervisé	Exécute les actions réversibles dans des limites définies.	L'humain contrôle par exception et traite les alertes.
N3 • Agent autonome cadré	Conduit un processus complet dans un périmètre autorisé.	L'humain audite, règle les politiques et prend les cas sensibles.
N4 • Système multi-agents	Coordonne plusieurs agents spécialisés et leurs outils.	L'humain gouverne les identités, les coûts, les droits et les résultats.

Autonomie opérationnelle croissante → · Supervision : de l'exécution directe à la gouvernance.

Figure 5 : Cinq niveaux d'autonomie. Le périmètre de l'agent s'élargit, tandis que la responsabilité humaine évolue de l'exécution directe vers la gouvernance.

NIVEAU	DESCRIPTION	SUPERVISION HUMAINE
N0 • Assistant	Répond à la demande, produit du texte.	Totale : l'humain exécute.
N1 • Copilote	Propose une action prête à valider.	Validation systématique.
N2 • Agent supervisé	Exécute des actions réversibles, escalade le reste.	Contrôle par exception.
N3 • Agent autonome cadré	Conduit un processus complet dans un périmètre défini.	Audit a posteriori, seuils d'alerte.
N4 • Système multi-agents	Plusieurs agents spécialisés coordonnés par un orchestrateur.	Gouvernance, observabilité, politiques.

Tableau 23 : Échelle d'autonomie. La très grande majorité des déploiements rentables de 2026 se situe aux niveaux N2 et N3.

Les données de terrain confirment cette prudence. McKinsey observe que les organisations les plus performantes définissent bien plus souvent des processus explicites de validation humaine : 65 % chez les hauts performeurs contre 23 % pour les autres. L'autonomie se gagne par paliers, adossée à la mesure.

Source : McKinsey, State of AI, novembre 2025 ; analyse CX Today, février 2026.

5.8 Le standard qui a débloqué l'écosystème : MCP

La connexion d'un agent aux outils de l'entreprise a longtemps constitué un travail d'intégration sur mesure, à refaire pour chaque couple modèle-outil. Le Model Context Protocol, publié en open source par Anthropic en novembre 2024, standardise cette connexion : chaque application implémente le protocole une fois, chaque outil l'expose une fois.

- Adoption par les acteurs concurrents en moins de dix-huit mois : OpenAI, Google, Microsoft, AWS, GitHub et de nombreux éditeurs.
- Transfert de la gouvernance du protocole à l'Agentic AI Foundation, fonds dirigé sous l'égide de la Linux Foundation, le 9 décembre 2025 : avec Anthropic, OpenAI et Block comme cofondateurs, AWS, Google, Microsoft, Cloudflare et Bloomberg comme membres platine.
- Volume d'écosystème : la Registry officielle recensait plus de 9 600 serveurs référencés fin mai 2026 ; Anthropic évoquait plus de 10 000 serveurs publics actifs dans sa mise à jour de décembre 2025.
- Adoption en entreprise : une enquête sectorielle 2026 situe à 41 % la part des organisations logicielles interrogées exploitant des serveurs MCP en production, de façon limitée ou étendue.

Sources : Wikipedia, Model Context Protocol ; Anthropic, mise à jour écosystème décembre 2025 ; Digital Applied, relevé Registry API du 24 mai 2026 ; Stacklok, 2026 software report.

POURQUOI CE POINT COMPTE POUR UNE DIRECTION GÉNÉRALE

Un agent bâti sur un standard ouvert et gouverné par une fondation neutre réduit le risque d'enfermement propriétaire. Ce critère devient un point d'examen légitime dans tout appel d'offres agentique.

06

Comprendre le RAG : Retrieval-Augmented Generation

Le modèle cesse de répondre de mémoire. Il répond sur pièces.

Le RAG (Retrieval-Augmented Generation, ou génération augmentée par la recherche) associe la puissance de formulation des modèles de langage aux sources de connaissance validées de l'organisation. Le système recherche d'abord les passages pertinents dans une base gouvernée, puis les transmet au modèle avec la question. Résultat : des réponses plus exactes, à jour et traçables.

LE RAG EN BREF — UNE ARCHITECTURE AUGMENTÉE, PAS MAGIQUE

Le RAG ne remplace pas le modèle de langage : il l'enrichit. Avant de produire une réponse, le système interroge une base documentaire pour récupérer les informations pertinentes, puis les injecte dans le contexte de génération. Cette architecture améliore la précision factuelle, réduit les hallucinations et rend exploitables les connaissances propres à l'entreprise : politiques internes, procédures, référentiels, questions fréquentes, bases métier.

6.1 Définition du RAG : Retrieval-Augmented Generation

Le RAG relie le modèle aux connaissances validées de l'organisation. Au lieu de demander au LLM de répondre uniquement à partir de ce qu'il a appris pendant son entraînement, le système recherche d'abord les passages pertinents dans une base documentaire gouvernée, puis les fournit au modèle avec la question.

Le RAG n'est donc ni un LLM supplémentaire ni un agent. C'est une méthode d'accès à la connaissance. Il apporte des faits, des sources et un contexte actualisable ; il ne décide pas seul du parcours à suivre, des outils à appeler ou du moment où il faut transmettre la conversation à un humain.

Le chapitre précédent a montré comment un agent agit. Encore faut-il qu'il agisse sur la base d'une information exacte : c'est le rôle du RAG. Interroger une base de connaissance au moment de la question, puis fournir au modèle les passages pertinents pour qu'il fonde sa réponse dessus : le modèle cesse de répondre de mémoire, il répond sur pièces.

6.2 Pourquoi ne pas simplement agrandir le contexte ?

Les fenêtres de contexte des modèles de langage se sont considérablement élargies : plusieurs centaines de milliers de jetons chez les principaux fournisseurs. Cette évolution alimente une objection fréquente à l'utilité même du RAG : puisque l'on peut désormais insérer une documentation entière dans la requête, l'ancrage documentaire deviendrait superflu. Trois réalités opérationnelles le maintiennent pourtant au centre des architectures de production.

- L'économie. Le coût d'inférence croît avec le nombre de jetons traités. Injecter systématiquement 300 000 jetons pour répondre à une question qui en mobilise 2 000 revient à multiplier la facture par un facteur considérable, à chaque interaction, en production.
- La latence. Un agent conversationnel qui répond en dix secondes perd l'interlocuteur. La recherche ciblée conserve un temps de réponse compatible avec une conversation en direct.
- La précision. La qualité de restitution se dégrade sur les segments situés au milieu d'un contexte très long. Une sélection préalable des passages pertinents améliore la fidélité de la réponse à la source.

Les faits de marché confirment cette lecture : les grandes plateformes ont continué d'investir massivement dans la recherche documentaire managée en 2026, au moment même où leurs fenêtres de contexte atteignaient des tailles inédites. AWS a ainsi lancé au Sommet de New York de juin 2026 des Knowledge Bases entièrement managées, dotées d'un « Agentic Retriever » destiné aux requêtes complexes en plusieurs étapes.

Source : AWS News Blog, annonces AWS Summit New York, juin 2026.

6.3 Le pipeline, étape par étape

SCHÉMA 6

Le pipeline RAG

Neuf étapes : cinq hors ligne pour construire la connaissance, quatre en ligne pour répondre.

PHASE HORS LIGNE — ÉTAPES 1 À 5 : CONSTRUIRE LA BASE DE CONNAISSANCE

1	INGESTION Collecte des sources : documents, site, CRM, catalogue, tickets, e-mails.
2	PARSING Extraction structurée du contenu : tableaux, PDF, images, mise en forme.
3	DÉCOUPAGE Segmentation en fragments cohérents et contextualisés.
4	VECTORISATION Conversion des fragments en représentations numériques.

5	INDEXATION Stockage en base vectorielle, doublé d'un index lexical.
---	---

BASE DE CONNAISSANCE — RAG + MÉMOIRE ET CONTEXTE

PHASE EN LIGNE — ÉTAPES 6 À 9 : RÉPONDRE À UNE QUESTION, AVEC SES SOURCES

6	RECHERCHE Récupération des passages pertinents pour la question posée.
7	RERANKING Reclassement fin des candidats par un modèle spécialisé.
8	GÉNÉRATION ANCRÉE Réponse rédigée à partir des seuls passages fournis, avec citations.
9	ÉVALUATION Mesure de la fidélité, de la pertinence et de la couverture.

| Boucle d'amélioration continue : l'évaluation renvoie vers les étapes hors ligne.

Figure 6 : Le pipeline RAG — une phase d'indexation hors ligne qui alimente une base de connaissance, puis une phase de réponse déclenchée à chaque question. Le détail de chaque étape figure au tableau 24, page suivante.

ÉTAPE	FONCTION	DÉCISIONS DÉTERMINANTES
1. Ingestion	Collecte des sources : documents, site, CRM, catalogue, tickets, e-mails, bases métier.	Périmètre, fraîcheur, droits d'accès, fréquence de synchronisation.
2. Parsing	Extraction structurée du contenu : tableaux, PDF, images, mise en forme.	Qualité de l'extraction des tableaux et des documents scannés.
3. Découpage	Segmentation en fragments cohérents et contextualisés.	Taille des fragments, recouvrement, conservation des titres parents.
4. Vectorisation	Conversion des fragments en représentations numériques (embeddings).	Choix du modèle d'embedding, dimension, langue, coût de réindexation.
5. Indexation	Stockage dans une base vectorielle, souvent doublée d'un index lexical.	Isolation par client, métadonnées, filtres de sécurité.
6. Recherche	Récupération des passages les plus pertinents pour la question posée.	Recherche hybride (sémantique et mots-clés), nombre de passages.
7. Reranking	Reclassement fin des candidats par un modèle spécialisé.	Gain de précision au prix d'une latence supplémentaire.
8. Génération ancrée	Rédaction de la réponse à partir des seuls passages fournis, avec citations.	Consigne d'abstention quand la source manque.
9. Évaluation	Mesure continue de la fidélité, de la pertinence et de la couverture.	Jeux de test métier, évaluateurs de modèle, revue humaine.

Tableau 24 : Les neuf étapes d'un pipeline RAG de production.

6.4 Ce que le RAG apporte, en chiffres

- Exactitude et fidélité. Les travaux académiques récents mesurent des gains nets. Une équipe de l'IIT Madras a montré, sur des tâches de questions-réponses financières, qu'un pipeline combinant RAG et orchestration agentique améliorait l'exactitude de 11,2 % et réduisait les hallucinations de 15 %.
- Réduction massive en contexte métier fermé. Un déploiement universitaire documenté (système MARAUS, plus de 6 000 interactions réelles) rapporte une exactitude moyenne de 92 % et un taux d'hallucination ramené de 15 % à 1,45 %, avec des temps de réponse inférieurs à quatre secondes.
- Traçabilité. Chaque affirmation renvoie à une source identifiable. Cette propriété conditionne l'usage dans les secteurs régulés — santé, finance, droit, secteur public — qui mènent l'adoption du RAG selon l'ensemble des cabinets d'analyse.
- Valorisation du patrimoine documentaire. Plus de 80 % des données d'entreprise sont non structurées ; le RAG est le mécanisme qui les rend interrogeables par une IA.

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.

LE RAG EN UNE PHRASE

Le RAG remplace « le modèle croit savoir » par « le modèle sait où regarder, et il cite ce qu'il a lu ».

6.5 Les limites du RAG employé seul

Un pipeline RAG classique exécute une séquence fixe : une question, une recherche, une réponse. Cette linéarité rencontre quatre plafonds.

1. Les questions composées. « Quel est le délai de livraison pour la référence X vers l'Espagne, et reste-t-il du stock ? » exige deux recherches distinctes dans deux sources distinctes.
2. Le raisonnement en plusieurs sauts. Une réponse qui dépend d'un fait établi par une première recherche demande une seconde recherche informée par la première.
3. L'absence d'autocritique. Un pipeline linéaire produit une réponse même quand les passages récupérés se révèlent hors sujet.
4. L'absence d'action. Le RAG répond avec exactitude ; la prise de rendez-vous, l'écriture au CRM et la relance restent en dehors de son périmètre.

Ces quatre plafonds recoupent exactement le périmètre de l'architecture agentique présentée au chapitre précédent : c'est cette complémentarité que le chapitre suivant met en système.

6.6 La base de connaissances : le premier actif IA durable de l'entreprise

Le point d'entrée le plus réaliste de l'IA en entreprise n'est pas toujours un projet de productivité. Il peut être plus simple, plus utile et plus consensuel : rendre la connaissance existante accessible, gouvernée et exploitable. Une organisation possède déjà l'essentiel de sa matière première dans ses pages web, ses offres, ses procédures, ses catalogues, ses contrats types, ses réponses de support, ses comptes rendus, son CRM et la mémoire de ses experts. Le problème tient à la dispersion, à l'obsolescence et aux droits d'accès.

Construire une base de connaissances pour le RAG ne signifie pas verser tous les fichiers dans un moteur de recherche. Il faut identifier la source qui fait autorité, le propriétaire éditorial, la date de validité, le public autorisé, la fréquence de mise à jour et la réponse attendue lorsqu'une information manque. Ce travail crée une couche de connaissance exploitable par plusieurs systèmes et plusieurs équipes. Même si le premier cas d'usage est commercial, l'actif produit peut ensuite servir au support, à l'intégration des nouveaux arrivants, aux opérations, à la formation ou à l'assistance interne.

AVANT	TRANSFORMATION	ACTIF OBTENU
Documents dispersés	Inventaire et classification	Corpus connu et attribué
Versions contradictoires	Source d'autorité et dates de validité	Réponse cohérente
Savoir détenu par quelques personnes	Formalisation et validation	Mémoire organisationnelle transmissible
Accès indifférencié	Métadonnées et contrôle d'accès	Recherche conforme aux habilitations
Mise à jour ponctuelle	Synchronisation et alertes	Connaissance vivante
Réponses non mesurées	Jeux de questions et évaluation	Qualité observable

Tableau 25 : Le projet RAG transforme un stock de fichiers en couche de connaissance gouvernée.

Cette couche constitue une base de développement IA pour toute la société parce qu'elle découple la connaissance du canal. Les mêmes faits validés peuvent alimenter un agent sur le site web, une conversation WhatsApp, un échange téléphonique, un échange par e-mail, une notification SMS, un

assistant interne ou un parcours d'après-vente. Les règles de formulation et d'action changent ; la source de vérité reste commune. L'entreprise évite ainsi de créer cinq assistants incohérents qui apprennent séparément cinq versions de la même offre.

PRINCIPE D'ARCHITECTURE

La connaissance doit être mutualisable sans devenir indifférenciée : un socle commun, des droits d'accès explicites, des vues adaptées à chaque métier et une isolation stricte des données propres à chaque client ou entité.

6.7 Une même connaissance sur le web, WhatsApp, le téléphone, l'e-mail et le SMS

Le RAG n'est pas lié à une interface. Il intervient après la compréhension de la demande et avant la réponse ou l'action. Sur le web, l'entrée est textuelle et peut être enrichie par la page consultée. Sur WhatsApp, la conversation est asynchrone, persistante et souvent faite de messages courts, de pièces jointes ou de notes vocales. Au téléphone, la parole doit être transcrite, traitée et restituée avec une latence très faible. Dans les trois cas, l'agent peut interroger la même base de connaissances, appliquer les mêmes règles et conserver un contexte relationnel unifié.

CANAL	CE QUE L'AGENT REÇOIT	ADAPTATION NÉCESSAIRE	VALEUR HUMAINE
Web	Texte, page, campagne, comportement de navigation	Réponse immédiate, liens, formulaires et prise de rendez-vous	Le commercial récupère une demande déjà contextualisée
WhatsApp	Messages, historique, documents, audio	Conversation asynchrone, consentement, reprise sans répétition	Le conseiller reprend le fil exact, au bon moment
Téléphone	Voix, rythme, interruptions, intention	Transcription et synthèse en temps réel, phrases courtes, escalade fluide	L'humain intervient sur l'émotion, la complexité ou la négociation
E-mail	Message long, pièces jointes, fil de discussion	Réponse structurée, citations, délai et validation éventuelle	L'expert valide les engagements sensibles
SMS	Message bref et identité mobile	Réponse concise, liens et rappels	Le rendez-vous et la continuité sont sécurisés

Tableau 26 : Le canal change la forme de l'interaction, pas la source de vérité.

La cohérence multicanale ne consiste donc pas à copier le même message partout. Une bonne architecture sépare trois couches : la connaissance, commune et validée ; le raisonnement, qui détermine l'intention et la prochaine étape ; l'expression, adaptée au canal. Une réponse téléphonique doit être courte et naturelle, tandis qu'un e-mail peut être plus détaillé et documenté. Ce qui ne doit pas varier, ce sont les conditions tarifaires, les règles d'éligibilité, les limites de promesse ou l'historique du client.

Ce modèle facilite l'adoption interne. L'entreprise peut commencer par un seul périmètre de connaissance et un seul canal à faible risque, mesurer la qualité, puis étendre progressivement. Chaque extension réutilise l'actif documentaire, les jeux d'évaluation, les règles et l'historique déjà construits. Le déploiement devient cumulatif plutôt qu'une succession de prototypes isolés.

6.8 Sécuriser la connaissance : une architecture gouvernée, pas un simple copier-coller

Le RAG ne rend pas une information sûre par le seul fait de l'indexer. La sécurité vient de l'architecture qui l'entoure. Une base de connaissances d'entreprise doit conserver les règles d'accès des sources, séparer les espaces de chaque organisation, filtrer la recherche selon l'identité et le rôle, chiffrer les données au repos et en transit, protéger les secrets des connecteurs, limiter la conservation et journaliser les consultations comme les actions. L'agent ne doit retrouver que ce que la personne et le cas d'usage sont autorisés à connaître.

RISQUE	CONTRÔLE ATTENDU DANS UNE ARCHITECTURE RAG AGENTIQUE
Fuite entre utilisateurs ou clients	Cloisonnement des espaces, contrôle d'accès par rôle et filtres appliqués avant la récupération
Document sensible retrouvé hors contexte	Métadonnées de sensibilité, droits hérités de la source et politiques d'exclusion
Injection ou document malveillant	Validation des sources, analyse des contenus, séparation entre instructions et documents récupérés
Action métier non autorisée	Permissions minimales, validation humaine selon le risque et liste d'outils explicitement autorisés
Absence de preuve	Journal des sources consultées, des réponses, des décisions, des outils appelés et des transferts humains
Conservation excessive	Durées de rétention définies, suppression contrôlée, chiffrement et choix contractuel du fournisseur de modèle

Tableau 27 : La sécurité du RAG dépend de contrôles explicites à chaque couche.

Le risque le plus courant n'est donc pas le nom d'un outil : c'est l'usage d'un assistant sans cadre d'entreprise. Copier une information confidentielle dans un compte personnel, sans politique interne, sans contrôle des réglages, sans gestion des droits et sans engagement contractuel sur les données, crée une exposition évitable. Les conditions diffèrent selon les offres. OpenAI indique par exemple que les services individuels peuvent utiliser le contenu pour améliorer les modèles sauf désactivation, tandis que ChatGPT Business, ChatGPT Enterprise et l'API n'utilisent pas par défaut les entrées et sorties de l'organisation pour entraîner les modèles.

Une architecture comme Sacha Engine doit donc choisir le modèle et son mode de connexion selon le niveau de sensibilité, la localisation, la politique de conservation et les engagements contractuels attendus. Le RAG dédié maintient la connaissance dans un périmètre gouverné ; l'orchestrateur applique les droits et les règles ; le LLM reçoit uniquement le contexte nécessaire à la réponse ; les actions sensibles exigent un contrôle supplémentaire. La protection de la connaissance devient ainsi une propriété vérifiable du système, et non une promesse générale.

PRINCIPE DE SÉCURITÉ

La bonne question n'est pas seulement « quel modèle utilisons-nous ? », mais « quelles données lui transmettons-nous, pour quel utilisateur, avec quels droits, pendant combien de temps et avec quelle trace ? »

Références : NIST AI 600-1, OWASP LLM08:2025 et OpenAI Business Data Privacy ; détails en annexe D.

07

IA agentique + RAG + LLM : pourquoi c'est la combinaison gagnante

L'agentique décide et agit. Le RAG apporte la connaissance de l'entreprise. Le LLM comprend et formule. Séparées, aucune ne suffit.

Prises séparément, les trois briques ne répondent chacune qu'à une partie du problème : un agent sans connaissance gouvernée agit sur une information insuffisante ; un RAG sans agent retrouve l'information sans jamais conduire un parcours ; un LLM seul formule bien, mais à partir de sa mémoire d'entraînement et non de vos données. Réunies, elles forment un système opérationnel — informé, traçable et maîtrisable — dont ce chapitre décrit l'assemblage, couche par couche.

Avant d'entrer dans l'assemblage, retenir ceci : ces trois briques sont aussi les trois questions à poser à n'importe quel fournisseur d'agent IA. Elles ne demandent aucune compétence technique, et les réponses vous disent en trois minutes à qui vous avez affaire.

LA QUESTION À POSER	CE QU'ELLE VÉRIFIE	LA RÉPONSE QUI DOIT VOUS ALERTER
« Quel modèle utilisez-vous, et puis-je en changer ? »	Le LLM : la capacité à comprendre et à formuler — et votre liberté de ne pas dépendre d'un seul fournisseur.	« C'est notre IA propriétaire », sans nommer le modèle ni prévoir d'en changer.
« D'où vient la réponse quand l'agent parle de mes prix, de mes offres ou de mes dossiers ? »	Le RAG : une connaissance de l'entreprise à jour, sourcée et vérifiable phrase par phrase.	« Le modèle a été entraîné sur vos données » ou « il apprend tout seul en discutant ».
« Que sait-il faire, et jusqu'où va-t-il sans moi ? »	L'agentique : la capacité d'agir — proposer un créneau, envoyer un document, relancer, transmettre — et les limites de cette autonomie.	« Il répond à toutes les questions », sans liste d'actions ni règle de passage à l'humain.

Tableau 28 : Trois questions, trois briques. Si l'une des trois manque, l'agent aura un défaut structurel : il inventera, il ne saura pas, ou il n'agira pas.

Ces trois questions ne disent pas comment le système est construit. Elles disent s'il possède les trois briques — c'est-à-dire s'il peut fonctionner du tout. La mise en œuvre est un autre sujet : outils et intégrations, mémoire, garde-fous, évaluation, relais humain. Elle se vérifie au moment du cahier des charges, en partie III.

7.1 Intégrer les LLM, le RAG et l’agentique dans l’entreprise

L’ordre pédagogique ne doit pas être confondu avec le parcours technique. Pour comprendre le système, on distingue d’abord l’agentique, puis le RAG, puis le LLM. En fonctionnement, ces briques dialoguent en boucle : une demande entre par un canal, l’agent identifie l’objectif et les règles, le RAG recherche les éléments autorisés, le LLM interprète et formule, puis l’agent transmet la réponse, appelle un outil autorisé ou passe le relais à l’humain.

ÉTAPE	BRIQUE PRINCIPALE	RÔLE DANS L'ENTREPRISE
01	Demande et canaux	Web, WhatsApp, téléphone, e-mail ou SMS : la personne exprime son besoin dans son langage et au moment où il apparaît.
02	Couche agentique	Elle comprend l'objectif, choisit les étapes, applique les règles, décide des outils autorisés et sait quand transmettre à l'humain.
03	RAG et base de connaissances	Ils recherchent les passages pertinents dans les contenus validés de l'entreprise, avec droits d'accès, versions et sources.
04	LLM	Il interprète la question et les éléments retrouvés, puis prépare une réponse claire, contextualisée et adaptée à la langue du canal.
05	Systèmes, humain et traçabilité	Le système diffuse, met à jour les outils autorisés, conserve le contexte et remet la demande à l'équipe compétente lorsque son intervention crée la valeur.

Tableau 29 : Les cinq étapes qui relient la demande entrante à l’action, et la brique responsable de chacune.

Cette séparation évite les confusions : un LLM seul sait formuler mais ne connaît pas nécessairement l’entreprise ; un RAG seul retrouve mais ne conduit pas un parcours ; une couche agentique sans connaissance gouvernée risque d’agir sur une information insuffisante. La valeur apparaît lorsque les trois sont reliés aux équipes, aux règles, aux outils et aux canaux réels de l’organisation.

ARCHITECTURE PUBLIQUE · SACHA OMON AI

Dans cette architecture, Sacha Engine constitue la couche d’orchestration propriétaire de SACHA OMON AI. Il relie les canaux, une couche d’intelligence sélectionnable, la connaissance gouvernée de l’entreprise, les actions autorisées et le relais humain. Le schéma public ci-dessous montre cette répartition des responsabilités sans détailler la logique interne de routage, de construction du contexte, de mémoire, de RAG correctif (CRAG), de contrôle ou d’exécution.

SCHÉMA 7

Architecture opératoire SACHA OMON AI

Sacha Engine relie les canaux, l'intelligence, la connaissance et le relais humain.

CANAUX	Web · WhatsApp · Téléphone · E-mail · SMS
SACHA ENGINE	Agent IA agentique · orchestration propriétaire, sécurisée et confidentielle
INTELLIGENCE ET CONNAISSANCE	LLM sélectionnable · RAG dédié
RÉPONSE ET RELAIS HUMAIN	Service continu · contexte transmis

Orchestration propriétaire · sécurisée · confidentielle.

Figure 7 : Architecture publique de SACHA OMON AI. Sacha Engine orchestre l'agent IA, la couche d'intelligence, la connaissance d'entreprise et le relais humain, sans exposer les mécanismes propriétaires.

7.2 Le RAG devient agentique : l'agent décide quand chercher et quand agir

DIMENSION	RAG CLASSIQUE	RAG AGENTIQUE
Déclenchement	Recherche systématique à chaque requête.	Recherche décidée par l'agent, selon le besoin réel.
Requête	La question de l'utilisateur, telle quelle.	Question décomposée et reformulée en plusieurs requêtes complémentaires.
Sources	Un index unique.	Routage vers la bonne source : documentaire, CRM, catalogue, base graphe, web.
Itération	Un seul passage.	Boucles successives, chaque résultat informant la recherche suivante.
Contrôle qualité	Génération directe.	Autocritique : suffisance des preuves évaluée avant réponse.
Issue	Une réponse rédigée.	Une réponse et une action exécutée dans les systèmes métiers.
Traçabilité	Citation des passages utilisés.	Journal complet : plan, appels d'outils, sources, décision, action.

Tableau 30 : Du pipeline linéaire à la boucle pilotée par le raisonnement.

7.3 Pourquoi l'agentique RAG améliore les résultats

1. Décomposition de requête. L'agent scinde une question composée en sous-questions autonomes, puis recompose. Les travaux de l'IIT Madras appliquent ce principe via la génération de requêtes multiples non équivalentes, et mesurent le gain conjoint sur l'exactitude et sur les hallucinations.

2. Routage vers la source pertinente. Une question tarifaire interroge le catalogue ; une question de statut interroge le système de gestion ; une question de méthode interroge la documentation. Le routage évite la dilution de la pertinence dans un index unique.
3. Recherche itérative multi-sauts. L'agent enchaîne les recherches en s'appuyant sur les faits déjà établis, ce que les jeux d'évaluation multi-sauts mesurent explicitement.
4. Recherche hybride et reclassement. La combinaison d'une recherche lexicale et d'une recherche sémantique, suivie d'un reclassement, constitue selon la littérature récente le socle de récupération le plus robuste ; le raisonnement agentique intervient ensuite pour l'expansion contrôlée des preuves et la synthèse.
5. Autocritique et abstention. L'agent évalue la couverture des preuves et se déclare en incapacité plutôt que d'improviser : comportement décisif pour la confiance en contexte commercial ou réglementé.
6. Mémoire persistante. L'agent conserve l'historique du contact et l'intègre à ses recherches ultérieures, ce qui répond directement au défaut identifié par le MIT : l'incapacité des outils génériques à retenir le contexte d'une session à l'autre.

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

7.4 Ce que la recherche établit, et ce qu'elle nuance

La synthèse académique de référence sur l'agentic RAG apporte deux enseignements utiles à une décision d'investissement.

- Le gain est le plus marqué dans les domaines à connaissance structurée et à contraintes explicites : santé, finance, droit. Les tâches ouvertes affichent des gains plus modestes. Autrement dit, la modélisation du domaine reste une composante à part entière du dispositif.
- Le coût du raisonnement supplémentaire mérite une mesure. Une étude comparative publiée en 2026 observe que la décomposition apporte un bénéfice dépendant du domaine et que les boucles de réflexion ajoutent une latence significative pour des gains de qualité inégaux. La conception doit donc calibrer le niveau d'itération sur la difficulté réelle de la requête.

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

LE PRINCIPE DE CONCEPTION À RETENIR

Un agent utile applique un effort proportionné : réponse directe sur les questions simples, recherche unique sur les questions factuelles, décomposition et itération sur les questions complexes. La sophistication uniforme dégrade la latence et le coût sans améliorer la qualité perçue.

7.5 Le schéma de référence

Application SACHA OMON AI : dans ce schéma, la couche « agent » correspond à l'agent IA SACHA OMON AI, orchestré par Sacha Engine.

SCHÉMA 8

Architecture de référence en cinq couches

Des canaux jusqu'à la gouvernance, qui traverse l'ensemble du dispositif.

1 · CANAUX — ENTRÉE ET SORTIE DES CONVERSATIONS	Site web · WhatsApp · Téléphone · E-mail · SMS
2 · ORCHESTRATION — RAISONNEMENT, PLANIFICATION, SÉLECTION D'OUTILS, RÉFLEXION	Agent : moteur de raisonnement agentique
3 · CONNAISSANCE ET MÉMOIRE — CE QUE L'AGENT SAIT ET RETIENT	Connaissance (RAG) : base vectorielle · index lexical · graphe · connecteurs métier — Mémoire : session · profil de contact · historique multicanal
4 · OUTILS — EXÉCUTION EFFECTIVE DES ACTIONS	CRM · Agenda · Facturation · Téléphonie · Paiement
5 · GOUVERNANCE — SÉCURITÉ, CONFORMITÉ, QUALITÉ, TRANSVERSE AUX QUATRE COUCHES	Journalisation · garde-fous · évaluation continue · escalade humaine

Figure 8 : Architecture de référence appliquée à l'agent IA SACHA OMON AI. La couche 3 réunit deux composants, la connaissance (RAG) et la mémoire, que le tableau 31 détaille séparément.

COUCHE	FONCTION	COMPOSANTS TYPES
Canaux	Entrée et sortie des conversations.	Site web, WhatsApp, téléphonie, e-mail, SMS.
Orchestration	Raisonnement, planification, sélection d'outils, réflexion.	Moteur agentique, politiques, machine à états.
Connaissance	Ancrage factuel de chaque réponse.	Base vectorielle, index lexical, graphe, connecteurs métier.
Mémoire	Continuité conversationnelle et relationnelle.	Mémoire de session, profil de contact, historique multicanal.
Outils	Exécution effective des actions.	CRM, agenda, facturation, téléphonie, e-mail, paiement.
Gouvernance	Sécurité, conformité, qualité.	Journalisation, garde-fous, évaluation continue, escalade humaine.

Tableau 31 : Architecture de référence d'un agent IA agentique augmenté par le RAG.

7.6 Du premier message à la réponse multicanale : le parcours concret

Le schéma public ci-dessous présente ce que vit l'utilisateur sans exposer les mécanismes propriétaires de Sacha Engine. Une personne engage une conversation sur le web, WhatsApp, le téléphone, l'e-mail ou le SMS. SACHA OMON AI mobilise l'intelligence nécessaire et la connaissance gouvernée de l'entreprise pour comprendre la demande, apporter une réponse utile, qualifier la situation et préparer la bonne suite sur le canal adapté.

L'entreprise bénéficie d'une expérience continue et d'une connaissance commune sur tous les canaux. Les mécanismes internes d'orchestration, de routage des modèles, de construction du contexte, de mémoire,

de contrôle, de CRAG et d'appel des outils constituent le savoir-faire propriétaire de Sacha Engine et ne sont volontairement pas détaillés dans ce livre blanc.

SCHÉMA 9

Du premier message à la réponse multicanale

Chaque couche échange avec les autres ; l'humain reste au centre du jugement, des validations et des décisions sensibles.

HUMAIN — JUGEMENT · NÉGOCIATION · DÉCISION · VALIDATION DES CAS SENSIBLES
· GOUVERNANCE ET RESPONSABILITÉ

1	PROSPECT OU CLIENT Web · WhatsApp · téléphone · e-mail · SMS. Question, demande, intention.
2	SACHA OMON AI Agent IA agentique · Sacha Engine. Comprend, raisonne, prépare la réponse et l'action adaptée.
3	CONNAISSANCE D'ENTREPRISE RAG dédié et gouverné. Faits, règles, contexte.
4	EXPÉRIENCE MULTICANALE Réponse · qualification · action.

Échanges bidirectionnels : l'humain orchestre et intervient au bon moment. Sacha traite et relie les échanges ; l'humain reçoit, valide, arbitre et réoriente lorsque sa valeur est nécessaire.

Figure 9 : Du premier message à la réponse multicanale. Les échanges sont bidirectionnels et l'humain reste au centre du jugement, des validations et de la gouvernance.

7.7 La couche d'intelligence : où se connectent les LLM

Dans l'architecture SACHA OMON AI, le LLM est une couche d'intelligence sélectionnable. Il comprend, raisonne et formule, mais il n'est ni la base de connaissances, ni la mémoire de l'entreprise, ni le détenteur des droits. Sacha Engine constitue la couche propriétaire d'orchestration ; le RAG dédié maintient la connaissance gouvernée de l'entreprise ; l'humain conserve le jugement, les exceptions et la responsabilité.

Cette séparation est stratégique : l'entreprise ne dépend pas d'un modèle unique et sa connaissance ne devient pas la propriété du LLM. Le modèle reçoit uniquement le contexte nécessaire à la tâche dans le cadre défini par l'architecture. Le choix du fournisseur, du niveau de raisonnement, de la latence et des contraintes de données peut évoluer sans déplacer l'actif central : la connaissance et les règles de l'entreprise.

Pour l'utilisateur, cette complexité reste invisible. Il pose sa question sur son canal ; SACHA OMON AI répond dans le même contexte, poursuit la conversation et transmet à un humain lorsque la situation

l'exige. Le livre blanc expose cette répartition des responsabilités pour comprendre l'architecture, sans publier la logique interne de routage, de contrôle, de mémoire ou d'exécution propre à Sacha Engine.

SCHÉMA 10

Quatre rôles complémentaires, une décision humaine

Comment l'IA travaille avec l'entreprise.

1 · LLM Comprend et formule Il interprète la demande et construit une réponse naturelle et utile.	2 · RAG Apporte les faits de l'entreprise Il retrouve la connaissance validée : offres, règles, documents et contexte.	3 · AGENTIQUE Conduit le parcours Il choisit l'action suivante, relance, qualifie et oriente au bon moment.	4 · HUMAIN Conserve la responsabilité Il exerce son jugement, valide les décisions sensibles et garde le dernier mot.
---	--	---	---

L'IA devient une capacité opérationnelle de l'entreprise. Elle répond là où la valeur se perd, tout en restant orchestrée par l'humain.

Figure 10 : Quatre rôles complémentaires, répartis sur les six composants techniques détaillés au tableau 32 — le LLM comprend et formule ; le RAG apporte les faits de l'entreprise ; l'agentique conduit le parcours ; l'humain conserve la responsabilité et le dernier mot.

COUCHE	À QUOI ELLE SERT	CE QU'ELLE NE DOIT PAS PRÉTENDRE ÊTRE
LLM	Comprendre, raisonner, planifier, synthétiser et formuler.	La vérité de l'entreprise, sa mémoire ou son système de droits.
Orchestrateur agentique	Choisir le modèle, router la demande, appliquer les politiques, gérer l'état et décider de l'étape suivante.	Un simple prompt envoyé une seule fois.
RAG dédié	Retrouver les passages autorisés, actuels et pertinents ; conserver le lien vers les sources.	Un entraînement du modèle ou un stockage indifférencié de documents.
Mémoire et contexte	Conserver l'identité, l'historique autorisé, l'état du dossier et la prochaine action.	Une mémoire implicite ou incontrôlée du fournisseur de modèle.
Outils métier	Créer un rendez-vous, lire un statut, enrichir un CRM, ouvrir un ticket ou préparer un paiement.	Une action simulée dans une réponse textuelle.
Gouvernance	Contrôler les identités, droits, validations, coûts, évaluations, traces et transferts humains.	Un contrôle ajouté après le déploiement.

Tableau 32 : Chaque couche a une responsabilité distincte. La robustesse vient de leur coordination, pas du LLM pris isolément.

À RETENIR

Le LLM est le moteur de compréhension et de raisonnement. Le RAG lui apporte les faits de l'entreprise. L'orchestrateur conduit le parcours. La mémoire conserve le contexte autorisé. Les outils produisent l'effet réel. La gouvernance contrôle l'ensemble.

7.8 Une architecture multi-modèles plutôt qu'une dépendance à un fournisseur

Sacha Engine peut connecter plusieurs familles de modèles : OpenAI GPT, Anthropic Claude, Google Gemini, xAI Grok, ainsi que des modèles ouverts comme Llama ou Mistral lorsque le contexte le justifie. L'objectif n'est pas d'accumuler les fournisseurs, mais d'éviter qu'un choix unique impose la même latence, le même coût, le même niveau de raisonnement et les mêmes contraintes de données à tous les cas d'usage.

La capacité multi-modèles permet d'adapter le niveau d'intelligence au contexte sans publier les règles internes de sélection. Une interaction en temps réel, une analyse complexe, une tâche multilingue ou une donnée soumise à des contraintes régionales n'imposent pas nécessairement le même modèle. Sacha Engine applique la politique définie pour l'entreprise et conserve une architecture indépendante du fournisseur.

Cette indépendance protège aussi l'actif central : la connaissance de l'entreprise. Les documents, métadonnées, droits, historiques et jeux d'évaluation ne doivent pas être enfermés dans un modèle. Ils restent dans la couche Sacha Engine et peuvent être mobilisés par le LLM le mieux adapté à la situation.

Sources : documentations officielles OpenAI (Agents), Anthropic (Claude et MCP), Google Cloud (Gemini et Vertex AI Agent Builder), xAI (Grok et outils agentiques), Meta (Llama) et Mistral AI.

B. DE LA DEMANDE À L'ACTION COMMERCIALE

L'architecture appliquée au cas d'usage où la valeur se perd le plus vite.

08

De la demande à l'action commerciale immédiate : l'agent SDR IA agentique multicanal

Entre le clic payé et la réponse apportée se trouve l'endroit exact où la valeur se perd.

Les premières pages de ce Livre blanc ont posé un changement de comportement : lorsque le besoin apparaît, nous nous habituons à obtenir une réponse tout de suite. Ce chapitre examine ce que cette attente change dans la relation commerciale — capter la demande au moment où elle existe, y répondre, la qualifier, la documenter, puis la transmettre à l'humain avec l'intégralité du contexte.

8.1 La demande n'attend plus : le temps réel devient la nouvelle règle commerciale

En 1996, Internet nous a appris à chercher l'information. En 2026, l'IA nous habitue de plus en plus à formuler directement un besoin et à recevoir immédiatement une information, une recommandation, un service, parfois même une action. Ce changement de mentalité ne reste pas confiné aux assistants personnels : il modifie progressivement ce que les personnes attendent aussi des entreprises.

Lorsqu'un prospect clique sur une publicité, ouvre WhatsApp, arrive sur un site ou appelle, son intention existe maintenant. Il ne raisonne pas en fonction des horaires du service commercial ou du délai moyen de rappel. La première exigence devient donc de commencer immédiatement la relation : comprendre pourquoi la personne est là, lui apporter une première réponse utile, préciser son besoin et préparer l'étape suivante.

Les chiffres confirment que la fonction commerciale a déjà commencé cette bascule. Selon State of Sales 2026 de Salesforce, conduit auprès de plus de 4 000 professionnels de la vente, 87 % des organisations commerciales utilisent déjà l'IA dans leur cycle de vente : prospection, prévision, notation des prospects ou

rédaction d'e-mails. La question n'est donc plus seulement de savoir si l'IA entre dans la vente, mais comment elle se relie au moment exact où la demande apparaît.

IBM précise le fonctionnement d'un agent SDR IA : dès qu'un prospect remplit un formulaire ou ouvre une conversation, l'agent peut poser les questions de qualification, interpréter les réponses et vérifier leur correspondance avec les critères définis par l'entreprise, sans attendre qu'un collaborateur soit disponible.

McKinsey apporte la précision essentielle pour une direction générale : la valeur ne vient pas d'un bouton IA ajouté à un parcours existant, mais d'un workflow repensé autour de l'agent. Son étude sur les workflows marketing agentiques évoque jusqu'aux deux tiers des activités marketing actuelles et une hausse de chiffre d'affaires de 10 à 30 % associée à l'hyperpersonnalisation.

Le cas Verizon illustre cette bascule à grande échelle. Selon Reuters, en avril 2025, l'assistant IA déployé auprès de 28 000 conseillers a réduit la durée des appels et libéré du temps commercial, avec une hausse des ventes proche de 40 % depuis son déploiement.

LE FIL ROUGE DU CHAPITRE 8

Rupture 1 : la demande attend une réponse immédiate. Rupture 3 : la provenance, le consentement et la preuve deviennent des données opérationnelles. L'agent SDR IA se situe exactement à l'intersection de ces deux mouvements.

Sources : Salesforce, State of Sales 2026, 7^e édition, plus de 4 000 répondants, février 2026 ; IBM Think, étude sur la redéfinition de la vente par les agents SDR IA, avril 2026 ; McKinsey, « Reinventing marketing workflows with agentic AI », 2026 ; Reuters, avril 2025.

APPLICATION : SACHA OMON AI

SACHA OMON AI applique directement ce principe : l'agent SDR IA agentique multicanal répond en moins de deux secondes, 24 heures sur 24, sur le web, WhatsApp, le téléphone, l'e-mail et le SMS, en s'appuyant sur une base de connaissance RAG dédiée à chaque entreprise cliente.

8.2 La preuve par le marché : cent millions de prospects jamais rappelés, et l'acquisition de Qualified par Salesforce

Un épisode survenu chez le plus grand éditeur de CRM au monde illustre, mieux qu'aucune étude, l'ampleur du problème que résout un agent SDR IA agentique.

100 millions

De leads restés sans réponse au cours d'environ un quart de siècle, selon Marc Benioff
The Logan Bartlett Show, 29 août 2025

10 000

Leads qu'un représentant agentique peut rappeler en une semaine
Exemple cité par TIME, 28 août 2025

68 000

Leads analysés et réengagés par Salesforce
Chiffre publié au 1^{er} octobre 2025

Marc Benioff a expliqué que Salesforce avait accumulé environ 100 millions de leads sans suivi, faute de capacité humaine suffisante. Salesforce a ensuite documenté l'utilisation d'Agentforce pour analyser, recontacter et qualifier ces leads. Au 1^{er} octobre 2025, l'entreprise indiquait 68 000 leads nouvellement analysés et réengagés, 800 rendez-vous réservés automatiquement et des dizaines de ventes conclues.

CE QUE CE CAS MONTRE

Une entreprise peut investir massivement dans l'acquisition et perdre malgré tout de la valeur après le clic si la demande n'est pas traitée. Le sujet n'est donc pas seulement de générer davantage de leads : il est d'être capable de répondre au moment où l'intention existe, puis de transmettre à l'humain les cas qui exigent son jugement.

Précision éditoriale : ce cas public est cité comme repère de marché. Il ne constitue ni un partenariat, ni une approbation de SACHA OMON AI par Salesforce ou Marc Benioff.

Le 29 août 2025, Marc Benioff, cofondateur, président et directeur général de Salesforce, a révélé, lors d'un entretien accordé au podcast The Logan Bartlett Show, que plus de 100 millions de prospects ayant contacté Salesforce au cours des 26 années précédentes n'avaient jamais été rappelés, faute de capacité humaine suffisante. Salesforce a depuis déployé sa propre technologie de vente agentique pour engager plus de 10 000 prospects entrants par semaine et transformer ces conversations en pipeline commercial.

Quatre mois plus tard, le 17 décembre 2025, Salesforce a annoncé la signature d'un accord définitif en vue d'acquérir Qualified, éditeur de solutions de marketing agentique et d'agents SDR IA dont le produit Piper qualifie et convertit en continu les visiteurs entrants d'un site web. L'acquisition a été finalisée le 1^{er} avril 2026. Dans son Form 10-Q pour le trimestre clos le 30 avril 2026, Salesforce indique que la juste valeur de la contrepartie transférée pour Qualified était de 1,2 milliard de dollars, principalement 1,1 milliard de dollars en numéraire.

Cette opération constitue un signal de marché à part entière : les technologies capables d'engager, de qualifier et de convertir la demande entrante sans rupture de continuité sont désormais valorisées à l'échelle du milliard de dollars, y compris par l'éditeur qui venait, en interne, de démontrer publiquement le coût de leur absence.

Sources : Marc Benioff, The Logan Bartlett Show, 29 août 2025 ; Salesforce, communiqué du 17 décembre 2025, mis à jour le 1^{er} avril 2026 ; Salesforce, Form 10-Q, trimestre clos le 30 avril 2026.

8.3 Les marchés commencent à intégrer le risque IA des centres de contacts

Les marchés financiers ont commencé à intégrer un risque devenu stratégique pour les grands opérateurs de centres de contacts : lorsque des agents IA peuvent comprendre une demande, consulter une base de connaissances, répondre sur le web ou par téléphone et transmettre le contexte à un conseiller, la valeur ne peut plus reposer uniquement sur le volume de personnes mobilisées.

Le 28 février 2024, l'action Teleperformance a perdu 27 % en une séance et touché un plus bas de sept ans, Reuters reliant explicitement ce mouvement aux inquiétudes suscitées par la capacité de l'IA à transformer le métier des centres d'appels. La figure de la page suivante compare l'évolution boursière sur deux ans de Teleperformance et de trois groupes comparables cotés aux États-Unis.

SCHÉMA 11

Marché boursier · opérateurs de relation client : la réévaluation sur deux ans

Clôtures non ajustées · comparaison du 13 août 2024 au 13 août 2026.

SOCIÉTÉ	CLÔTURE AU 13 AOÛT 2024	DERNIÈRE CLÔTURE	ÉVOLUTION
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 %

Lecture : sur deux ans, les quatre groupes ont perdu entre 33,0 % et 63,7 % de leur valeur boursière. Dernière clôture au 13 août 2026, et au 12 août pour le Nasdaq : au relevé du 13 août la séance était en cours, la dernière clôture complète a donc été retenue afin d'éviter tout cours intraday.

Figure 11 : Évolution boursière comparée sur deux ans de Teleperformance et de trois groupes de centres de contacts cotés aux États-Unis. Sources : Yahoo Finance (historiques TEP, CNXC, TTEC, TASK) · Euronext Paris · TTEC Investor Relations et LSEG · TaskUs Investor Relations et LSEG · calcul SACHA OMON AI.

Ces cours ne prouvent pas que l'IA explique seule la baisse des valorisations. La croissance organique, les marges, la dette, le mix clients et la qualité d'exécution restent déterminants. Ils montrent toutefois que les investisseurs réévaluent les modèles où la capacité de service dépend principalement du nombre de conseillers disponibles. Teleperformance ne reste d'ailleurs pas immobile : le groupe, qui indique opérer dans près de 100 pays, 400 langues et 170 marchés avec près de 490 000 collaborateurs, développe lui-même une offre TP.AI et une logique d'orchestration entre IA et humains.

LE SIGNAL À RETENIR

La relation humaine ne disparaît pas. Ce qui se fragilise est le modèle dans lequel la valeur repose seulement sur le nombre de conseillers disponibles. L'avenir appartient aux architectures qui rendent la connaissance immédiatement accessible, étendent la capacité de service et réservent à l'humain les situations où son jugement crée la différence.

Source du signal 2024 : Reuters, 28 février 2024. Sources des cours et méthodologie : indiquées dans la figure ci-contre.

8.4 Ce que l'humain ne peut pas faire à l'échelle : répondre à tout, en permanence

La productivité n'est pas le premier bénéfice de l'agentique en matière commerciale. Le premier bénéfice est structurel : un agent IA accomplit ce qu'aucune organisation humaine ne peut économiquement reproduire, à savoir traiter l'intégralité des demandes entrantes, à toute heure, sans file d'attente ni jour de repos. Cette capacité change la nature de l'expérience proposée au client, bien au-delà d'un simple gain de temps.

Le décalage se mesure d'abord au niveau du budget publicitaire. EMARKETER prévoit pour 2026 environ 243 milliards de dollars de recettes publicitaires numériques mondiales pour Meta et 240 milliards pour Google ; sur le seul marché américain, la publicité numérique est estimée à 413 milliards de dollars sur l'année. Ces prévisions d'analystes ne se comparent pas directement aux revenus publicitaires publiés par les deux groupes dans leurs comptes officiels (page 4), dont le périmètre est plus large. Ce qu'elles montrent, en revanche, ne varie pas : chaque dollar investi dans ces campagnes vise à générer une demande entrante.

Or une part significative de cette demande arrive en dehors des heures d'ouverture : HubSpot situe à 52 % la part des sollicitations téléphoniques reçues hors des horaires standards, un ordre de grandeur que plusieurs études sectorielles 2026 retrouvent pour les formulaires web et les demandes de rendez-vous.

Une étude portant sur 22 500 concessionnaires automobiles américains, menée par Foureyes et relayée par Ringlead en 2026, chiffre précisément ce que ce décalage coûte : pour un établissement dépensant en moyenne 45 000 dollars par mois en publicité et générant environ 150 demandes entrantes, 43 % ne reçoivent jamais d'appel réel, soit près de 19 350 dollars de budget publicitaire mensuel qui produisent des prospects qu'aucune voix humaine ne contacte jamais. Le même travail chiffre à environ 95 400 dollars par mois l'écart de marge entre un taux de réponse moyen et un taux de réponse inférieur à soixante secondes.

Sources : EMARKETER, prévisions publicité numérique, avril 2026 ; ResearchAndMarkets, United States Digital Ad Spend Market, février 2026 ; HubSpot ; Foureyes et Ringlead, étude sur 22 500 concessionnaires automobiles, 2026.

Cette demande non traitée ne se heurte pas seulement à un problème d'horaires. Les formulaires web imposent un délai et une friction que l'acheteur d'aujourd'hui tolère de moins en moins ; la génération précédente de chatbots, construite sur des arbres de décision figés, butait sur toute question sortant du script prévu : c'est précisément la limite que les chapitres 5 et 6 ont détaillée. Un agent SDR IA agentique lève ces deux obstacles à la fois : il répond avec précision, sur la base d'une connaissance ancrée par RAG, puis oriente la conversation selon son objet — rendez-vous avec un décideur, mise en relation avec un commercial, ou traitement direct d'une demande de service client — plutôt que de laisser la personne sans réponse, en attente d'un rappel qui n'arrivera pas.

CE QUE L'AGENTIQUE CHANGE STRUCTURELLEMENT

Un centre d'appels humain, même opéré 24 heures sur 24, reste borné par le nombre de conseillers disponibles à un instant donné et par l'étendue de leurs compétences face aux objections. Un agent IA agentique traite un nombre de conversations simultanées qui n'est pas contraint par un effectif, pour un coût par interaction sans commune mesure avec le coût chargé d'un conseiller humain. Le paragraphe 8.9 détaille cette économie.

8.5 Le canal devient une partie de la vente : web, WhatsApp et voix

Le web, WhatsApp et le téléphone sont trois portes d'entrée différentes vers le même besoin. Sur le web, la personne arrive depuis une recherche, une campagne ou une page précise et attend une réponse pendant que son intention est encore active. Une publicité Meta peut aussi ouvrir directement une conversation WhatsApp : Meta décrit les annonces « click-to-WhatsApp » comme des publicités Facebook ou Instagram qui envoient la personne directement dans une discussion avec l'entreprise.

WhatsApp est devenu un canal commercial majeur. Meta indique que plus de 3 milliards de personnes utilisent WhatsApp dans plus de 180 pays, et que plus de 2 milliards l'utilisent chaque jour. En juin 2026, Meta indiquait également plus d'un milliard de conversations actives avec des entreprises chaque jour sur WhatsApp, Messenger et Instagram, et plus d'un million d'entreprises utilisant déjà Meta Business Agent sur WhatsApp et Messenger. Pour l'entreprise, l'intérêt est simple : le clic peut devenir immédiatement une conversation, sans imposer un formulaire intermédiaire.

La voix complète ce dispositif. La WhatsApp Business Calling API permet notamment de recevoir des appels initiés par les utilisateurs. Dans cette logique, l'enjeu principal reste l'entrant : une personne clique, écrit ou appelle parce que son besoin existe maintenant. L'agent IA doit comprendre cette demande, mobiliser la connaissance de l'entreprise et répondre avant de transmettre à l'humain si nécessaire. L'e-mail et le SMS jouent ensuite un rôle complémentaire pour confirmer, envoyer un document, un lien, un rendez-vous ou assurer le suivi.

Sources officielles : Meta, « Click to Message Ads » et « Create Ads that Click to WhatsApp » ; Meta, « Ways To Manage Your Businesses Chats On WhatsApp », avril 2025 ; Meta, « Be There for Every Customer With Meta Business Agent », juin 2026 ; Meta for Developers, WhatsApp Business Calling API.

8.6 La troisième rupture devient opérationnelle : consentement, provenance et demande entrante

La troisième rupture présentée dans l'ouverture prend ici une forme opérationnelle. La provenance de la demande, le consentement, le canal utilisé et la capacité à en conserver la preuve deviennent des éléments de l'architecture commerciale, et pas seulement des sujets juridiques traités à part.

En France, la loi n° 2025-594 du 30 juin 2025 fait basculer, à compter du 11 août 2026, le régime du démarchage téléphonique d'un système d'opposition — la liste Bloctel — vers un système de consentement préalable. Le consentement doit être libre, spécifique, éclairé, univoque et révoquant ; il n'est valable qu'un an au maximum, et sa preuve doit être conservée par le professionnel durant trois ans au moins. Le décret d'application n° 2026-662 du 23 juillet 2026 en précise les modalités de recueil, de conservation et de retrait. Le non-respect des dispositions des articles L. 223-1 à L. 223-5 expose à une amende administrative pouvant atteindre 75 000 € pour une personne physique et 375 000 € pour une personne morale. En outre, l'article L. 223-1 prévoit que tout contrat conclu à la suite d'un démarchage téléphonique réalisé en violation de ses dispositions est nul. Séparément, lorsqu'un abus de faiblesse ou d'ignorance est caractérisé dans les conditions prévues par l'article L. 132-14-1, les peines sont portées à cinq ans d'emprisonnement et 500 000 € d'amende ; le montant de l'amende peut être porté à 10 % du chiffre d'affaires moyen annuel. Bloctel cesse d'exister à cette même date.

À l'échelle de l'Union européenne, la situation appelle une lecture précise plutôt qu'une extrapolation directe depuis le cas français. La directive ePrivacy (2002/58/CE) impose déjà, dans les vingt-sept États membres, un consentement préalable pour la prospection par courrier électronique, SMS et automates d'appel. Pour les appels téléphoniques passés par un être humain en revanche, cette même directive laisse à chaque État membre le choix entre un régime d'opposition et un régime de consentement préalable. La proposition de règlement ePrivacy, qui aurait pu harmoniser ce point à l'échelle européenne, a été formellement retirée par la Commission européenne le 16 juillet 2025, faute d'accord entre colégislateurs. La bascule française relève donc d'un choix national, dans un mouvement que plusieurs pays européens ont déjà engagé à des degrés divers, sans qu'un règlement unique et contraignant ne figure aujourd'hui à l'agenda de Bruxelles.

Aux États-Unis, la trajectoire récente va en sens inverse. Fin 2023, la Federal Communications Commission avait adopté une règle dite « one-to-one consent » : le consentement devait être donné à un vendeur nommément désigné, et non partagé entre plusieurs partenaires commerciaux. L'esprit en était proche de la réforme française.

Cette règle n'a pas survécu. La cour d'appel fédérale du onzième circuit a annulé ces restrictions le 24 janvier 2025, jugeant que la FCC avait excédé son autorité. La FCC a ensuite supprimé ces restrictions de son règlement par une règle finale publiée au Federal Register le 29 août 2025.

Le cadre applicable reste donc le régime antérieur : le consentement écrit préalable demeure requis pour les robocalls et robotexts publicitaires ou de télémarketing relevant du TCPA, notamment lorsqu'ils utilisent un système de composition automatique ou une voix artificielle ou préenregistrée. Le National Do Not Call Registry continue en parallèle de s'appliquer aux sollicitations de télémarketing relevant de son champ. Ce dispositif est structurellement plus proche de l'ancien régime français Bloctel que du nouveau régime d'opt-in généralisé.

La conséquence stratégique demeure la même des deux côtés de l'Atlantique, quoique pour des raisons distinctes. En France et chez plusieurs de ses voisins européens, la contrainte réglementaire réduit mécaniquement le volume de prospection sortante praticable, ce qui déplace la valeur vers la capacité à capter et convertir la demande entrante. Aux États-Unis, où cette contrainte ne s'est pas resserrée, l'argument reste principalement concurrentiel : la vitesse de réponse détermine déjà la conversion, comme le montrent les données des paragraphes 8.4 et 8.8, indépendamment de toute évolution réglementaire à venir.

Sources : Légifrance, loi n° 2025-594 du 30 juin 2025 et décret n° 2026-662 du 23 juillet 2026 ; economie.gouv.fr, DGCCRF ; Commission européenne, retrait de la proposition de règlement ePrivacy, 16 juillet 2025 ; Insurance Marketing Coalition v. FCC, 11^e circuit, 24 janvier 2025 ; FCC, règle finale, septembre 2025.

DE L'ANALYSE À L'APPLICATION

Ce qui précède décrit le marché. Ce qui suit décrit comment SACHA OMON AI y répond.

8.7 L'anatomie d'un agent SDR IA agentique multicanal

Un agent SDR IA agentique multicanal applique l'architecture décrite au chapitre 7 au cycle commercial amont. Le raisonnement agentique conduit la conversation et l'action ; le RAG garantit que chaque affirmation faite au prospect provient de la connaissance validée de l'entreprise.

SCHÉMA 12

Le parcours client d'un agent SDR IA agentique multicanal

Sept étapes ; à chacune, ce que conduit le raisonnement agentique et ce qu'apporte le RAG.

ÉTAPE	CE QUE FAIT L'AGENT IA	CE QU'APPORTE LE RAG
1 • Captation	Repère l'intérêt, le besoin initial et le canal.	Ouverture du contact sur le bon canal et avec le bon langage.
2 • Détection d'intention	Détecte le projet, le niveau d'urgence et l'intention.	Validation du produit, du besoin et du cas d'usage.
3 • Qualification	Qualifie la demande, les critères et le profil.	Critères d'éligibilité, cas, freins et zones.
4 • Réponse aux objections	Répond aux freins, rassure et reformule.	Arguments validés et réponses fondées sur la connaissance.
5 • Prise de rendez-vous	Propose un créneau, confirme et collecte les informations.	Règles d'attribution, agenda et priorités.
6 • Transmission	Crée le dossier et transmet au commercial.	Compte rendu structuré, contexte transmis au CRM.
7 • Relance	Relance si besoin et maintient l'engagement.	Règles de relance, scénarios et cadence contrôlée.

Figure 12 : Le parcours client d'un agent SDR IA agentique multicanal, étape par étape.

ÉTAPE DU CYCLE	APPORT DU RAISONNEMENT AGENTIQUE	APPORT DU RAG
Captation	Engagement immédiat sur le canal choisi par le visiteur, à toute heure.	Réponse d'ouverture contextualisée par l'offre et la page consultée.
Détection d'intention	Lecture du signal d'achat, adaptation du parcours de questions.	Compréhension du vocabulaire produit et des cas d'usage documentés.
Qualification	Séquence de questions adaptative, plutôt qu'un formulaire fixe.	Critères d'éligibilité, contraintes tarifaires et géographiques issues des sources internes.
Réponse aux objections	Reformulation, relance ciblée, gestion du doute.	Argumentaire, références, garanties et conditions, cités depuis la source.
Prise de rendez-vous	Consultation d'agenda, proposition, confirmation, rappel.	Règles d'affectation par territoire, expertise et disponibilité.
Transmission	Création de l'enregistrement CRM, transfert humain sur seuil défini.	Compte rendu structuré et sourcé de la conversation.
Relance	Cadence de suivi calibrée, sur le canal le plus engageant.	Historique complet du contact, tous canaux confondus.

Tableau 33 : Répartition des rôles entre raisonnement agentique et connaissance ancrée sur le cycle commercial amont.

8.8 Les indicateurs à instrumenter

INDICATEUR	DÉFINITION	REPÈRE DE MARCHÉ
Délai de première réponse	Écart entre l'arrivée de la demande et la première réponse utile.	Cible de projet : sous 60 secondes. Repère historique : dans l'audit publié par Harvard Business Review en 2011 (hbr.org/2011/03/the-short-life-of-online-sales-leads) portant sur 2 241 entreprises américaines, le délai moyen était de 42 heures parmi les entreprises ayant répondu dans les 30 jours.
Taux de captation hors heures ouvrées	Part des demandes traitées en dehors des plages humaines.	Gisement structurel : environ 75 % du temps calendaire.
Taux de qualification autonome	Part des conversations qualifiées sans intervention humaine.	À mesurer sur le périmètre réel du projet ; aucun benchmark transversal unique n'est retenu.
Taux de rendez-vous tenus	Rendez-vous honorés rapportés aux rendez-vous posés.	Indicateur de qualité de qualification, plus fiable que le volume.
Fidélité aux sources	Part des réponses intégralement appuyées sur la base de connaissance.	Mesurée en continu par évaluateur de modèle et revue humaine.
Taux d'escalade	Part des conversations transférées à un humain.	Un taux stable et assumé vaut mieux qu'un taux nul.
Coût par rendez-vous qualifié	Coût complet rapporté au rendez-vous obtenu.	Suivi par canal, type de conversation et niveau d'autonomie. L'objectif est de mesurer la performance du dispositif, pas de produire une comparaison artificielle avec un poste humain.

Tableau 34 : Tableau de bord minimal d'un agent SDR IA agentique multicanal.

8.9 L'économie d'un agent : mesurer le coût réel, sans plafond artificiel

Il n'existe pas de prix universel ni de plafond pertinent pour un agent IA. Le coût dépend du modèle choisi, des volumes de tokens, de la voix, des canaux, du RAG, des outils appelés, des intégrations, de l'infrastructure, du niveau de supervision et du degré d'autonomie confié au système.

La bonne lecture économique consiste donc à suivre les coûts réels du dispositif et à les relier à des résultats utiles : conversation traitée, demande qualifiée, rendez-vous tenu, résolution, action accomplie ou chiffre d'affaires attribuable. Cette mesure doit évoluer avec l'architecture, car les modèles, les plateformes et les méthodes de développement changent rapidement.

PRINCIPE

Mesurer ce que l'agent coûte et ce qu'il produit, sans prétendre qu'une fourchette de prix ou une comparaison homme-machine résume l'économie d'un projet.

APPLICATION : SACHA OMON AI

SACHA OMON AI est un agent SDR IA agentique multicanal : un agent IA unique qui agit dès la demande entrante sur le site web, WhatsApp, le téléphone, l'e-mail et le SMS, avec un contexte centralisé sur l'ensemble de ces canaux, 24 heures sur 24 et dans plus de 40 langues.

Son architecture applique le motif décrit dans ce livre blanc : le moteur agentique Sacha Engine conduit le raisonnement, la détection d'intention d'achat et la logique de qualification ; une base de connaissance RAG dédiée ancre chaque réponse dans la documentation validée du client ; une infrastructure mono-locataire isole les données de chaque client, avec hébergement européen ou américain ; les opérations connectées relient l'agent IA au CRM, à l'agenda, au transfert humain et au tableau de bord.

L'agent IA SDR est opéré, maintenu et mis à jour par SACHA OMON AI : le client mobilise ainsi ses équipes sur la conversion des rendez-vous, plutôt que sur l'exploitation d'une plateforme technique.

Cette architecture interopérable — RAG dédié, protocoles ouverts plutôt que propriétaires — positionne également l'entreprise pour l'étape suivante du marché, celle où l'agent personnel du client dialoguera directement avec l'agent de l'entreprise (paragraphe 2.5).

8.10 Du token au coût unitaire d'une tâche agentique

Le token constitue l'une des unités économiques centrales des LLM : les fournisseurs facturent généralement les volumes d'entrée et de sortie, avec des modalités qui varient selon les modèles, le cache, le contexte ou le niveau de raisonnement. Dans un agent IA, une seule demande peut mobiliser plusieurs appels au modèle ; la maîtrise des tokens devient donc un enjeu de conception et d'exploitation.

Pour autant, le coût d'un agent ne se réduit pas aux tokens. Il faut additionner, selon le cas, la recherche et le RAG, les appels d'outils, le calcul, le stockage, la voix ou la téléphonie, les canaux et la supervision. L'indicateur pertinent pour le dirigeant est le coût complet par résultat utile : conversation qualifiée, demande résolue, rendez-vous préparé ou action accomplie, comparé à la valeur métier produite.

Cette économie explique pourquoi le calibrage est essentiel. Un agent bien conçu utilise le niveau de raisonnement nécessaire, évite les boucles inutiles et réserve les modèles les plus coûteux aux situations qui le justifient. Le coût de production numérique reste alors à mettre en regard du temps humain économisé, de la disponibilité ajoutée, de la qualité de service et du chiffre d'affaires protégé ou créé.

8.11 Une intelligence, cinq portes d'entrée

Un agent commercial ou de service ne devrait pas devenir une collection de bots séparés. Le prospect qui commence sur le web, poursuit sur WhatsApp, reçoit un e-mail ou un SMS puis appelle ne vit qu'une seule relation. L'architecture doit donc unifier l'identité, l'historique, les règles et la prochaine action, tout en adaptant l'expression au canal.

Le parcours suit six étapes constantes. Le connecteur de canal transforme le message, la voix ou la pièce jointe en événement exploitable. Le système rattache cet événement au bon contact et récupère le contexte autorisé. Le moteur agentique identifie l'intention et décide s'il doit répondre, rechercher, demander une précision ou agir. Le RAG fournit les passages validés. L'outil métier exécute l'action

éventuelle. Enfin, la réponse est reformulée pour le canal et la trace est conservée pour le prochain échange ou le transfert humain.

ÉTAPE	QUESTION TRAITÉE	EXEMPLE
1. Normaliser	Que vient-il de se passer sur ce canal ?	Transcrire un appel ou lire une pièce jointe WhatsApp
2. Identifier	Qui parle et quel contexte peut être utilisé ?	Rattacher le numéro au bon contact CRM
3. Raisonner	Quelle est l'intention et la prochaine étape utile ?	Distinguer une question, une objection et une demande urgente
4. Ancrer	Quelles sources autorisées répondent à la demande ?	Récupérer le tarif, la règle et la date de validité
5. Agir	Quel effet doit être produit dans les systèmes ?	Créer une opportunité ou réserver un créneau
6. Transmettre	Comment continuer sans perdre le contexte ?	Envoyer la synthèse, les sources et la prochaine action au conseiller

Tableau 35 : Le même cycle agentic RAG derrière chaque canal.

Sur WhatsApp, l'agentique est particulièrement utile parce que la conversation est discontinuée. Un prospect peut répondre plusieurs heures plus tard, envoyer une photo, une note vocale ou une simple confirmation. La mémoire externe conserve l'état du parcours ; le RAG retrouve l'information exacte ; l'agent décide s'il peut continuer ou s'il doit transférer. La messagerie devient ainsi un espace de service durable, et non un chatbot de session.

Sur le web, la vitesse et le contexte de navigation dominant. La page consultée, la campagne d'origine et le comportement observé peuvent aider à choisir les premières questions, sous réserve du cadre de consentement. Le RAG évite que l'agent improvise une offre ; l'intégration à l'agenda ou au CRM permet de transformer immédiatement une intention en étape commerciale.

Au téléphone, la contrainte principale est la latence. La chaîne voix, transcription, raisonnement, recherche puis synthèse vocale doit répondre en temps réel, interrompre proprement et éviter les formulations trop longues. La connaissance reste la même, mais le moteur d'expression sélectionne des réponses brèves. Dès qu'une émotion forte, une négociation complexe ou un risque apparaît, l'appel peut être transféré avec la transcription, le motif et les éléments déjà vérifiés.

8.12 Avant-vente et après-vente : deux portes d'entrée complémentaires

L'avant-vente est un point d'entrée naturel parce qu'elle combine forte valeur et tâches répétables : comprendre le besoin, répondre aux questions, comparer des options, traiter les objections simples, vérifier une éligibilité, recueillir les informations utiles et préparer un rendez-vous. L'agent ne remplace pas le commercial ; il lui remet une demande comprise, documentée et déjà qualifiée, au moment où l'intervention humaine apporte le plus de valeur.

L'après-vente est tout aussi important. Une base de connaissances RAG peut réunir les notices, procédures, références de matériels, consignes d'entretien, codes d'erreur, questions fréquentes et règles de garantie. Le client obtient ainsi une première réponse utile sans attendre l'ouverture du service ou répéter son problème à plusieurs interlocuteurs.

Sur WhatsApp, le cas d'usage devient très concret : le client peut envoyer la photo d'un équipement, d'un écran de contrôle, d'un code erreur, d'une pièce ou d'un schéma. Un modèle capable d'analyser l'image peut décrire ce qu'il voit ; le RAG retrouve ensuite la notice ou la procédure validée correspondante ; l'agent formule les premières vérifications autorisées et conserve le contexte de l'échange.

Pour l'entretien d'une pompe à chaleur, par exemple, l'agent peut aider à identifier un code affiché, rappeler une opération d'entretien documentée, vérifier un réglage prévu par le fabricant ou préparer un diagnostic de premier niveau. Si la situation est ambiguë, si une action présente un risque ou si une intervention physique est nécessaire, le système alerte et transmet directement au technicien la photo, l'historique, les vérifications déjà réalisées et les sources consultées.

Le bénéfice n'est donc pas de supprimer le service après-vente. Il est de filtrer les demandes simples, de donner une réponse immédiate quand elle est sûre et documentée, puis de réserver le temps des techniciens aux cas qui exigent réellement leur expertise. La même connaissance peut servir avant la vente, pendant l'installation et plusieurs années après la signature.

MOMENT	CE QUE L'AGENT PRÉPARE	CE QUE L'HUMAIN PORTE
Découverte	Questions initiales, contexte, documentation pertinente	Écoute stratégique et reformulation complexe
Qualification	Critères, cohérence, disponibilité, synthèse	Arbitrage des exceptions et potentiel relationnel
Proposition	Données, options, calculs et pièces	Négociation, engagement et créativité commerciale
Intégration du client	Rappels, procédures, collecte structurée	Accompagnement du changement et relation
Support	Diagnostic de premier niveau, sources, ticket	Expertise, empathie et résolution des cas non standards
Fidélisation	Signaux, historique, actions suggérées	Compréhension politique et développement du compte

Tableau 36 : Une répartition des rôles conçue pour augmenter la capacité de service.

8.13 L'agent IA prend la disponibilité. L'humain garde le jugement et la relation.

Les échecs les plus visibles de l'IA en relation client ont souvent commencé par une promesse de substitution. Cette promesse produit des périmètres trop larges, des transferts trop tardifs et une mesure focalisée sur la réduction d'effectif. Une stratégie de complémentarité part de la question inverse : quelles capacités l'organisation veut-elle offrir qu'elle ne peut pas assurer aujourd'hui de manière constante ?

Ces capacités peuvent être une présence 24 heures sur 24, une réponse initiale dans toutes les langues importantes, une cohérence parfaite sur les conditions commerciales, une préparation systématique des dossiers, une relance consentie ou une transmission sans perte de contexte. L'humain reste responsable du jugement, des exceptions, de l'empathie, de la négociation et de la relation de confiance. L'agent assume la disponibilité, la mémoire, la recherche, la répétition fiable et l'exécution cadrée.

Ce cadrage facilite aussi l'acceptation par les équipes. Le système n'est plus présenté comme un observateur destiné à mesurer ou remplacer leur activité, mais comme une infrastructure de service qui élimine les angles morts et leur remet des situations mieux préparées. Les collaborateurs deviennent alors les principaux contributeurs à la qualité de la base de connaissances, aux règles d'escalade et aux jeux d'évaluation.

INDICATEUR DIRECTEUR

Mesurer d'abord la capacité de service ajoutée : demandes couvertes hors horaires, cohérence des réponses, qualité du contexte transmis, résolution au premier contact, satisfaction et conversion. Le temps économisé reste utile, mais il ne doit pas être l'unique récit de valeur.

8.14 Application : l'architecture SACHA OMON AI

SCHÉMA 13

Architecture fonctionnelle de SACHA OMON AI

Chaque couche interagit avec les autres ; la gouvernance s'applique transversalement à toute l'architecture.

1 · CANAUX — ENTRÉE ET SORTIE DES CONVERSATIONS	Site web · WhatsApp · Téléphone · E-mail · SMS. Échanges bidirectionnels : l'agent interagit avec les utilisateurs sur tous les canaux.
2 · ORCHESTRATION — RAISONNEMENT, PLANIFICATION, SÉLECTION D'OUTILS, RÉFLEXION	Agent : moteur de raisonnement agentique. Comprend l'intention · planifie · choisit l'action · pilote le parcours.
3 · CONNAISSANCE ET MÉMOIRE — CE QUE L'AGENT SAIT ET RETIENT	Connaissance (RAG) : base vectorielle · index lexical · graphe · connecteurs métier. Mémoire : session · profil de contact · historique multicanal. L'agent interroge pour comprendre, puis enrichit et met à jour.
4 · OUTILS — EXÉCUTION EFFECTIVE DES ACTIONS	CRM · Agenda · Facturation · Téléphonie · Paiement. L'agent sélectionne et active les outils ; il en reçoit résultats, statuts et données de retour.
5 · GOUVERNANCE — SÉCURITÉ, CONFORMITÉ, QUALITÉ, TRANSVERSE AUX QUATRE COUCHES	Journalisation · garde-fous · évaluation continue · escalade humaine. L'humain garde la responsabilité, le jugement et la décision.

Gouvernance transverse : elle supervise, protège et améliore en continu l'ensemble de l'architecture. Le paiement illustre une capacité future et non une fonction actuellement proposée.

Figure 13 : Architecture fonctionnelle publique de SACHA OMON AI. Les mécanismes internes de Sacha Engine restent confidentiels.

SACHA OMON AI applique ce modèle sous la forme d'un agent SDR IA agentique multicanal, déployé et opéré pour le client. Sacha Engine assure l'orchestration propriétaire et peut mobiliser plusieurs familles de LLM selon la politique de l'entreprise. Les modèles apportent compréhension, raisonnement et formulation ; la connaissance, les comptes et les actifs opérationnels restent gouvernés dans le périmètre du client.

Le livre blanc ne détaille volontairement ni la logique de routage, ni la construction du contexte, ni les mécanismes de mémoire, de CRAG, de contrôle ou d'exécution propres à Sacha Engine. Le principe public est simple : SACHA OMON AI comprend la demande, mobilise la connaissance autorisée, prépare l'action utile et transmet l'intégralité du contexte à l'humain lorsque son intervention est nécessaire.

Le service fonctionne sur le web, WhatsApp, le téléphone, l'e-mail et le SMS, 24 heures sur 24, dans plus de 40 langues. Sur les canaux temps réel, l'objectif de conception est une première réponse utile en moins de deux secondes lorsque les conditions techniques du canal le permettent. Le client n'a pas à exploiter lui-même une plateforme technique : SACHA OMON AI configure, déploie, opère, maintient et améliore le dispositif.

ARCHITECTURE DE RÉFÉRENCE

Infrastructure mono-locataire : données hébergées chez le client, exploitation managée et couche d'orchestration dédiée. Le différentiel ne vient pas du multicanal seul, mais de sa combinaison avec un RAG dédié, un contexte centralisé, une personnalisation métier profonde et une exploitation réalisée pour le client.

8.15 Ce que l'entreprise doit conserver comme actifs

Un projet agentique bien conçu ne se résume pas à un service consommé. Il doit laisser à l'entreprise des actifs réutilisables et portables. Cette exigence protège l'investissement, facilite le changement de fournisseur ou de modèle et permet d'étendre l'IA à de nouveaux métiers sans recommencer depuis zéro.

ACTIF	POURQUOI IL COMPTE	CONTRÔLE ATTENDU
Base de connaissances	Capitalise les sources, règles et validations	Exportable, documentée, propriétaires identifiés
Comptes et canaux	Conserve les numéros, identités et audiences	Ouverts et détenus au nom du client
Données relationnelles	Préserve l'historique et la continuité client	Stockées selon la politique et les droits du client
Règles et garde-fous	Formalise le fonctionnement métier	Lisibles, versionnés et auditables
Jeux d'évaluation	Permet de mesurer chaque évolution	Réutilisables quel que soit le modèle
Traces d'exécution	Explique les sources, outils, actions et escalades	Consultables avec une durée de conservation définie

Tableau 37 : Les actifs qui doivent rester utilisables par l'entreprise au-delà d'un fournisseur ou d'un modèle.

PARTIE III

MODE D'EMPLOI

A. Cadrer votre projet — sept questions structurantes, un script autonome, une méthode et quatorze décisions.

Ce livre défend l'idée qu'une entreprise doit rester maîtresse de sa connaissance, de ses données et de ses comptes. Il aurait été incohérent de vous vendre une méthode que vous ne pourriez appliquer qu'en dépendant de nous. Ce chapitre montre donc comment cadrer votre projet et rédiger une première version de votre cahier des charges avec votre propre IA, avant toute autre étape.

Sept questions structurantes. Votre point de départ.

Un projet d'agent IA ne commence pas par le choix d'un modèle ni par une comparaison de plateformes. Il commence par sept réponses structurantes. Elles déterminent le périmètre, la connaissance à mobiliser, les canaux, les règles, le dimensionnement et le mode de construction. Le script autonome présenté pages 88 à 92 ajoute une question d'ouverture — « En une phrase, avec vos mots, que voulez-vous que cet agent fasse ? » — avant de reprendre ces sept questions. Le script pose donc huit questions au total.

1	Quels marchés cet agent IA va-t-il servir ? Décide les langues, les canaux dominants, les règles locales et l'architecture régionale.
2	Dans quel secteur opérez-vous ? Décide les obligations réglementaires, le vocabulaire métier et les sources qui font autorité.
3	Quel est le premier usage visé ? Décide le périmètre, le résultat mesurable attendu et ce que l'agent doit transférer à un humain.
4	Où vit aujourd'hui la connaissance dont l'agent IA aura besoin ? Décide le travail de base de connaissances, les droits d'accès et la configuration du RAG.
5	Combien de demandes de ce type traitez-vous par mois ? Décide le dimensionnement, le coût réel du résultat utile et le retour attendu.
6	Quelle est la taille de votre organisation ? Décide la gouvernance, les niveaux de validation et le rythme de déploiement.
7	Qui va le construire ? Décide le partage des rôles : équipe interne, intégrateur, ou service opéré de bout en bout.

Répondez à ces sept questions et vous disposez du socle de votre projet.

CE QUE VOTRE PROPRE IA FAIT DE VOS RÉPONSES

De vos réponses au cahier des charges avec votre propre IA

Ces sept questions structurent le projet. Pour passer de ces réponses à un document exploitable, le script autonome des pages 88 à 92 ajoute une question d'ouverture — « En une phrase, avec vos mots, que voulez-vous que cet agent fasse ? » — puis reprend les sept questions précédentes. Il pose donc huit questions au total, une par une, dans votre propre IA — ChatGPT, Claude, Gemini ou Grok — avant de rédiger une première version du cahier des charges.

Aucun accès extérieur n'est nécessaire : vous copiez le script contenu dans ce livre blanc et vous le collez dans l'IA que vous utilisez déjà. Le document obtenu reste une hypothèse de travail destinée à être corrigée et validée par vos équipes.

VOUS OBTENEZ	CE QUE C'EST	CE QUE VOUS EN FAITES
Les 14 décisions, pré-rédigées	Une première rédaction des 14 décisions du cahier des charges, adaptée à vos réponses	Vous corrigez avec vos équipes au lieu de partir de zéro
La méthode appliquée à votre cas	Les 8 marches de déploiement, ordonnancées selon votre premier usage	Vous savez par quoi commencer et dans quel ordre
Une architecture cible	Les briques nécessaires et les points techniques à valider	Vous cadrez la discussion avec votre DSI ou votre prestataire
Les hypothèses de coût	Les ordres de grandeur du coût du résultat utile, pas seulement du modèle	Vous préparez un budget défendable
Les règles à vérifier	Les points réglementaires applicables à vos marchés et les questions à poser aux fournisseurs	Vous sécurisez le projet avant de l'engager

Tableau 38 : Ce que le script autonome produit à partir de vos huit réponses.

Aucune de ces cinq sorties n'est un engagement contractuel. Ce sont des hypothèses de travail destinées à être discutées, corrigées et validées par l'entreprise.

CE QUI SE PASSE ENSUITE · MÉTHODE SACHA OMON AI

Le déploiement en escalier : ne construisez pas tout, construisez la première marche

Une fois le cadrage posé, reste la question du rythme. La méthode défendue dans ce livre blanc est pragmatique : penser l'architecture globale dès le départ, mais ne pas figer pendant des mois toutes les briques d'un système. Les modèles, les plateformes et les outils évoluent trop vite pour cela. Chaque marche produit une preuve et prépare la suivante.

01

Besoin réel

Choisir un problème concret, un résultat attendu et un périmètre borné.

02

Preuve sur le terrain

Mettre une première capacité face à des cas réels et mesurer ce qui fonctionne.

03

Extension

Corriger la bonne couche, intégrer les progrès technologiques et ajouter la marche suivante.

Les huit marches opérationnelles

1. Choisir un besoin réel et mesurable.
2. Définir les connaissances et les règles qui font autorité.
3. Choisir un premier canal et un périmètre borné.

4. Donner à l'agent IA une première capacité utile.
5. Le confronter rapidement à des cas réels, sous contrôle.
6. Mesurer qualité, délai, actions, escalades et coût réel.
7. Corriger la bonne couche : données, règles, outil, modèle ou processus.
8. Ajouter la marche suivante en intégrant les progrès technologiques apparus entre-temps.

CONVICTION DU FONDATEUR

L'objectif n'est pas de construire aujourd'hui l'agent IA de 2028. Il est de construire aujourd'hui la première marche d'une architecture capable d'y arriver.

Cette trajectoire est la méthode proposée par Salomon Bellaïche dans ce livre blanc. Elle n'est pas une norme universelle : le bon rythme dépend du risque, du métier, des données et des responsabilités de chaque organisation.

LA GRILLE DE CONTRÔLE · LES 14 DÉCISIONS

Les 14 décisions à documenter — 1 à 7

Voici la grille complète. Le script autonome permet à votre propre IA d'en pré-rédiger une première version à partir de vos huit réponses. Cette grille sert ensuite de contrôle : tant qu'une décision n'a pas de preuve en face, elle n'est pas prise.

DOMAINE	QUESTION PRAGMATIQUE	PREUVE ATTENDUE
Objectif	Quel résultat concret l'agent IA doit-il obtenir ?	Résultat mesurable et responsable identifié
Périmètre	Quelles demandes traite-t-il, et lesquelles transfère-t-il ?	Cas inclus et exclus, règles d'escalade
Connaissance	Quelles sources font autorité ?	Inventaire, propriétaires, dates de mise à jour
Droits	Que peut-il voir et pour qui ?	Contrôles d'accès jusqu'à la recherche
Canaux	Où commence réellement la demande ?	Web, WhatsApp, téléphone, e-mail, SMS : volumes et contraintes
Mémoire	Quel contexte doit rester d'un échange à l'autre ?	Politique de conservation et identité du contact
Outils	Quelles actions l'agent IA peut-il réellement exécuter ?	Liste blanche, validation, réversibilité et journalisation

Tableau 39 : Les sept premières décisions du cahier des charges d'un agent IA.

VOUS POUVEZ LES RÉDIGER VOUS-MÊME, OU LES FAIRE PRÉ-RÉDIGER PAR L'IA

Le script autonome transforme vos huit réponses en une première hypothèse pour chacune de ces 14 décisions. Votre travail n'est ensuite pas de repartir de zéro, mais de corriger avec vos équipes.

MODE D'EMPLOI · LES 14 DÉCISIONS**Les 14 décisions à documenter — 8 à 14**

DOMAINE	QUESTION PRAGMATIQUE	PREUVE ATTENDUE
Humain	Quand l'agent IA doit-il passer la main ?	Seuils, routage et contexte transmis
Qualité	Comment sait-on qu'il répond et agit correctement ?	Cas réels, seuils, évaluations régulières
Sécurité	Comment limiter les injections, fuites et actions non autorisées ?	Isolation, filtres, tests et traces
Portabilité	Qui possède les comptes, numéros, données et règles ?	Propriété, export et réversibilité
Économie	Quel est le coût réel du résultat utile ?	Modèles, voix, outils, infrastructure, exploitation et supervision
Déploiement	Qui configure, opère, corrige et fait évoluer ?	Responsabilités, niveaux de service et processus de changement
International	Quelles langues, régions et réglementations sont concernées ?	Architecture régionale, localisation et règles par marché

Tableau 40 : Les sept dernières décisions du cahier des charges d'un agent IA.

Les pages 87 et 88 constituent la grille de contrôle du cahier des charges. Le script autonome permet à votre propre IA de préparer une première hypothèse à partir de vos huit réponses ; ensuite, le rôle de l'entreprise est de vérifier les sources, les droits, les actions, la sécurité, l'économie et les responsabilités humaines.

DE LA GRILLE AU CAHIER DES CHARGES

Vous avez maintenant une méthode en 8 marches et une grille de 14 décisions. Le script autonome présenté ci-dessous transforme ce cadrage en première version de cahier des charges avec votre propre IA. Le chapitre 9 commencera ensuite page 93 : « Comment construire et déployer un agent IA dans l'entreprise ».

Comprendre → Commencer → Construire.

MODE D'EMPLOI · CAHIER DES CHARGES AUTONOME**Faites rédiger votre cahier des charges par votre propre IA**

Le principe est simple. Vous copiez le script encadré ci-après et vous le collez dans votre propre intelligence artificielle — ChatGPT, Claude, Gemini ou Grok, celle que vous utilisez déjà. Elle vous pose alors huit questions, une par une, et attend votre réponse à chaque fois avant de poser la suivante.

Au terme de cet entretien, elle rédige votre cahier des charges : les 14 décisions documentées, la question finale, les 8 marches de déploiement situées sur votre cas, une première estimation du coût réel du

résultat utile, et les 5 questions à poser à tout fournisseur du chapitre 9.4. Dix à quinze minutes suffisent, sans préparation et sans abonnement supplémentaire : le script s'appuie sur l'IA que vous possédez déjà.

Ce livre défend l'idée qu'une entreprise doit rester maîtresse de sa connaissance, de ses données et de ses comptes. Il aurait été incohérent de vous vendre une méthode que vous ne pouvez appliquer qu'en dépendant de nous.

Le script commence ci-dessous et se poursuit sur les pages suivantes. Copiez-le intégralement, sans en changer un mot, puis collez-le dans une nouvelle conversation avec votre IA.

LE SCRIPT · À COPIER MOT POUR MOT

Tu es un chef de projet expérimenté en agents IA agentiques augmentés par le RAG. Tu appliques la méthode SACHA OMON AI, publiée dans le livre blanc « L'agent IA entre dans l'entreprise ».

TON RÔLE

Conduire un entretien court, puis rédiger un cahier des charges. Tu n'es ni un vendeur ni un enthousiaste. Tu es exigeant, concret, et tu dis quand quelque chose ne peut pas être décidé.

DÉROULÉ OBLIGATOIRE

Pose les huit questions ci-dessous UNE PAR UNE. Après chaque question, arrête-toi et attends ma réponse. Ne pose jamais deux questions à la fois. Ne rédige rien avant la huitième réponse. Si une réponse est vague, pose une seule relance, puis avance.

LES HUIT QUESTIONS

1. En une phrase, avec vos mots, que voulez-vous que cet agent fasse ?
2. Dans quels pays ou régions se trouvent vos utilisateurs ?
3. Dans quel secteur opérez-vous ?
4. Quel est le premier usage visé ? Un seul.
5. Où vit aujourd'hui la connaissance dont l'agent aura besoin ? (outils SaaS, fichiers dispersés, base structurée, papier ou PDF, dans la tête des équipes, un peu de tout)
6. Combien de demandes de ce type traitez-vous par mois ?
7. Quelle est la taille de votre organisation ?
8. Qui va le construire ? (équipe interne, agence ou intégrateur, éditeur clé en main, mélange, pas encore décidé)

APRÈS LA HUITIÈME RÉPONSE, PRODUIS LE DOCUMENT SUIVANT

A. LES 14 DÉCISIONS À DOCUMENTER

Pour chacune, écris quatre blocs dans cet ordre :

- EN CLAIR — une phrase simple, compréhensible par un dirigeant non technique.
- NOTRE PROPOSITION — un paragraphe rédigé pour MON cas, avec mes mots.
- PREUVE ATTENDUE — ce que je devrai pouvoir montrer.
- POURQUOI C'EST DÉCISIF — une phrase, sans emphase.

1. Objectif — Quel résultat concret l'agent doit-il obtenir ?

Preuve : résultat mesurable et responsable identifié.

2. Périmètre — Quelles demandes traite-t-il, et lesquelles transfère-t-il ?

Preuve : cas inclus/exclus et règles d'escalade.

3. Connaissance — Quelles sources font autorité ?

Preuve : inventaire, propriétaires, dates de mise à jour.

4. Droits — Que peut-il voir et pour qui ?

Preuve : contrôles d'accès appliqués jusqu'à la recherche.

5. Canaux — Où commence réellement la demande ?

Preuve : web, WhatsApp, téléphone, e-mail, SMS — volumes et contraintes.

6. Mémoire — Quel contexte doit rester d'un échange à l'autre ?

Preuve : politique de conservation et identité du contact.

7. Outils — Quelles actions l'agent peut-il réellement exécuter ?

Preuve : liste blanche, validation, réversibilité, journalisation.

8. Humain — Quand l'agent doit-il passer la main ?

Preuve : seuils, routage, contexte transmis.

9. Qualité — Comment sait-on qu'il répond et agit correctement ?

Preuve : cas réels, seuils, évaluations régulières.

10. Sécurité — Comment limiter les injections, fuites et actions non autorisées ?

Preuve : isolation, filtres, tests, traces.

11. Portabilité — Qui possède les comptes, numéros, données et règles ?

Preuve : propriété, export, réversibilité.

12. Économie — Quel est le coût réel du résultat utile ?

Preuve : modèles, voix, outils, infrastructure, exploitation, supervision.

13. Déploiement — Qui configure, opère, corrige et fait évoluer ?

Preuve : responsabilités, niveaux de service, processus de changement.

14. International — Quelles langues, régions et réglementations sont concernées ?

Preuve : architecture régionale, localisation, règles par marché.

B. LE RELEVÉ DE DÉPART

Demande-moi de relever trois chiffres avant de commencer, et explique en une phrase pourquoi : sans eux, personne ne pourra dire dans six mois si le projet a réussi. Délai moyen actuel ; volume traité par mois ; coût mensuel de cette activité ; date du relevé.

C. LA QUESTION FINALE

« Si l'agent cesse de fonctionner demain, que reste-t-il à l'entreprise ? »

Réponds pour mon cas. La bonne réponse comprend : une base de connaissances propre, des données exportables, des règles documentées, des jeux d'évaluation, des comptes sous contrôle et un processus humain renforcé. Dis-moi lesquels je n'ai pas.

D. LES HUIT MARCHES OPÉRATIONNELLES

Vision d'architecte, déploiement en escalier. Situe chaque marche sur MON cas :

1. Choisir un besoin réel et mesurable.
2. Définir les connaissances et les règles qui font autorité.
3. Choisir un premier canal et un périmètre borné.
4. Donner à l'agent une première capacité utile.
5. Le confronter rapidement à des cas réels, sous contrôle.
6. Mesurer qualité, délai, actions, escalades et coût réel.
7. Corriger la bonne couche : données, règles, outil, modèle ou processus.
8. Ajouter la marche suivante en intégrant les progrès apparus entre-temps.

E. LE COÛT RÉEL DU RÉSULTAT UTILE

À partir de mes réponses, estime uniquement ce qui peut raisonnablement l'être. Donne les ordres de grandeur du coût du modèle par résultat utile et du temps humain potentiellement redéployé. Pour la part de demandes que l'agent pourrait résoudre seul en régime établi, ne donne aucun pourcentage si mes réponses ne permettent pas de connaître suffisamment la nature et la complexité des demandes, les exceptions, l'état de la base de connaissances et les actions nécessaires. Dans ce cas, indique précisément les données qui manquent. Si tu proposes une fourchette, présente-la explicitement comme une hypothèse et explique sur quoi elle repose. Affiche toutes tes hypothèses. Rappelle les trois postes que les fournisseurs omettent souvent : la préparation initiale de la base de connaissances, la charge d'exploitation mensuelle et le coût des demandes qui resteront traitées par les équipes.

F. LES CINQ QUESTIONS À POSER À TOUT FOURNISSEUR

Pour chacune : la question mot pour mot, la réponse qui doit alerter, la réponse qui convient.

1. Sur quelles sources exactes votre agent s'appuie-t-il, et comment le vérifie-t-on ?
2. Quelles actions exécute-t-il vraiment, lesquelles exigent une validation humaine, et où est le journal ?

3. Comment mesurez-vous la qualité en production, et qui possède le jeu d'évaluation ?
4. À quel nom sont ouverts les comptes, les clés et les numéros ? Que puis-je exporter, dans quel format, sous quel délai ?
5. De combien de temps interne avez-vous besoin, de la part de qui, chaque mois ?

G. CE QUE TU NE PEUX PAS TRANCHER

Termine en listant les décisions qui ne peuvent pas être arrêtées à partir de huit réponses, et dis pourquoi : elles dépendent de mes contrats réels, de l'état réel de mes documents, de mon annuaire et de mes chiffres. Sois précis et honnête. Ne comble pas les trous.

RÈGLES ABSOLUES

- N'invente jamais une règle de droit, une date d'entrée en vigueur, un article ou un texte. Pour la décision 14, indique seulement les familles d'obligations à vérifier selon mes pays (protection des données, transparence sur l'usage de l'IA, règles sectorielles, contrôle à l'export si concerné), et écris que les dates et le détail doivent être vérifiés auprès d'un juriste ou d'une source officielle à jour.
- Ne dis jamais « automatisation ». Une automatisation exécute une règle écrite à l'avance ; un agent agentique décide dans un cadre.
- Écris dans la langue de la personne à qui tu t'adresses (français, anglais, hébreu, allemand, italien ou autre selon la langue qu'elle utilise), sobrement, sans emphase, sans superlatif, sans emoji.
- Ne présente aucun chiffre comme un engagement : ce sont des ordres de grandeur, et tu le dis.
- Si je te demande de sauter des étapes, refuse une fois et explique ce qui sera perdu.

COMMENCE MAINTENANT par la question 1, seule.

AVANT DE COMMENCER

Trois précautions

1. Le document obtenu est une base de travail, pas un document définitif : il se corrige avec vos équipes.
2. Les chiffres produits sont des ordres de grandeur, jamais des engagements.
3. Sur la réglementation, le script interdit à l'IA d'affirmer une date ou un texte : elle indique seulement les familles d'obligations à vérifier auprès d'un juriste.

Le chapitre suivant montre comment construire et déployer ce système dans l'entreprise.

B. DÉPLOYER DANS L'ENTREPRISE

Construire, gouverner, éviter les erreurs — puis conclure.

09

Comment construire et déployer un agent IA dans l'entreprise

*Penser l'architecture globalement. La
déployer marche par marche.*

Une intégration réussie ne commence pas par la remontée de tous les processus, de toutes les données et de tous les systèmes de l'entreprise : ce chantier retarde le résultat, dilue la responsabilité et nourrit les pilotes qui ne sont jamais adoptés. Elle choisit un point d'accès précis, là où l'entreprise supporte déjà le plus de valeur perdue — souvent l'acquisition marketing et le traitement de la demande entrante. Elle produit une preuve visible sur trafic réel, puis étend le périmètre une source, un canal, un outil et une équipe à la fois, jusqu'à constituer un centre nerveux commun.

9.1 Les sept décisions structurantes

DÉCISION	QUESTION À TRANCHER	RECOMMANDATION ISSUE DES RETOURS DE TERRAIN
Périmètre	Quel processus, quel volume, quel résultat mesurable ?	Un processus à fort volume et à faible variance, avec un indicateur financier identifié dès le cadrage.
Sources de connaissance	Quels documents, quelle fraîcheur, quels droits ?	Un périmètre restreint et parfaitement maintenu surpasse un périmètre large et approximatif.
Modèle	Quel moteur de raisonnement, quelle latence, quel coût ?	Modèles différenciés selon la difficulté de la requête ; portabilité conservée par principe.
Outils et intégrations	Quels systèmes l'agent écrit-il réellement ?	Standardiser sur un protocole ouvert ; commencer par la lecture, ouvrir l'écriture par paliers.
Autonomie	Où passe la frontière entre agent et humain ?	Actions réversibles confiées à l'agent, actions irréversibles soumises à validation.
Évaluation	Comment la qualité est-elle mesurée en continu ?	Jeu de test métier constitué avant le développement, exécuté à chaque évolution.
Propriété	À qui appartiennent les comptes, les données, les numéros ?	Comptes et données au nom du client, sans dépendance créée par le dispositif.

Tableau 41 : Les sept décisions à trancher avant la première ligne de code.

9.2 Commencez par un besoin précis, prouvez la valeur, puis élargissez

SCHÉMA 14

Commencer là où la valeur se perd déjà

Quatre critères pour choisir le point d'accès, puis six étapes de l'analyse à la centralisation.

AVANT DE COMMENCER — QUATRE CRITÈRES POUR CHOISIR LE POINT D'ACCÈS	
A	ACQUISITION Une dépense marketing déjà engagée.
B	DEMANDE ENTRANTE Des prospects et des clients qui attendent.
C	UN SEUL INDICATEUR Réponse, qualification ou revenu — pas les trois.
D	PAS DE TRANSFORMATION GLOBALE Un résultat que l'équipe peut voir.

LA TRAJECTOIRE — SIX ÉTAPES, DE L'ANALYSE À LA CENTRALISATION

01	ANALYSE CIBLÉE · J1 → J3 Un parcours, une perte, un responsable, un indicateur et les seules sources nécessaires.
02	PREMIER PÉRIMÈTRE · J4 → J7 Base RAG minimale validée, règles, escalade et un canal ou une campagne témoin.
03	TRAFIC RÉEL · J8 → J14 Mise en ligne contrôlée, journalisation, reprise humaine et correction quotidienne.
04	PREUVE · J15 → J30 Mesurer le délai, la qualité, l'adoption, la conversion, le service et le coût.
05	EXTENSION PROGRESSIVE · APRÈS LA PREUVE Une source, un canal, un outil, une équipe à la fois.
06	CENTRALISATION · EN CONTINU Connaissance, contexte, droits, traces et évaluations réunis dans Sacha Engine.

Principe de sortie : chaque extension doit conserver le gain mesuré, la qualité des réponses et la capacité de reprise humaine.

Figure 14 : Trajectoire recommandée — quatre critères pour choisir où commencer, puis six étapes détaillées au tableau 42, page suivante.

PHASE	DURÉE	LIVRABLES	CRITÈRE DE SORTIE
Analyse ciblée	J1 → J3	Un point d'accès, une perte, un responsable, un indicateur et les sources strictement nécessaires.	Périmètre, propriétaire et seuil de succès validés.
Premier périmètre	J4 → J7	Base RAG minimale validée, règles, escalade et un canal ou une campagne témoin.	Jeu de cas métier réussi sous contrôle humain.
Trafic réel	J8 → J14	Mise en ligne contrôlée, journalisation, reprise humaine et correction quotidienne.	Demandes réelles traitées de bout en bout.
Preuve	J15 → J30	Mesure du délai, de la qualité, de l'adoption, de la conversion, du service et du coût.	Gain visible et décision explicite de poursuivre.
Extension progressive	Après preuve	Ajout d'une source, d'un canal, d'un outil ou d'une équipe à la fois.	Chaque extension conserve l'indicateur et la qualité.
Centralisation	Continue	Connaissance, contexte, droits, traces et évaluations réunis dans Sacha Engine.	Centre nerveux commun, gouverné et réversible.

Tableau 42 : Mode de progression rapide. Les fenêtres J1-J30 supposent que les accès, les sources et les validations métier sont disponibles ; elles décrivent un premier point d'accès, jamais un déploiement global.

9.3 La règle de la donnée propre

Un constat traverse la totalité des retours d'expérience examinés : la préparation des données consomme plus de temps que le déploiement de l'agent lui-même. Une enquête sectorielle 2026 place la qualité des données au premier rang des freins au déploiement, citée par 52 % des répondants. Les études de cas les plus détaillées mentionnent explicitement des phases de nettoyage préalables plus longues que l'installation.

La conséquence pratique mérite d'être posée en tête de projet : le budget de préparation documentaire est un investissement de performance, et sa sous-estimation constitue la première cause d'écart entre le résultat annoncé et le résultat obtenu.

Source : compilations d'enquêtes entreprises 2026 ; études de cas éditeurs.

9.4 Les cinq questions à poser à tout fournisseur

1. Sur quelles sources l'agent fonde-t-il ses réponses, et comment le vérifiez-vous ? La réponse attendue décrit un dispositif de citation et une mesure de fidélité, plutôt qu'une affirmation de qualité.
2. Quelles actions l'agent exécute-t-il réellement dans nos systèmes ? L'écart entre « répondre » et « écrire dans le CRM et poser un rendez-vous » détermine la valeur produite.
3. Comment la qualité est-elle mesurée sur notre trafic après la mise en production ? L'évaluation continue sur données réelles constitue le meilleur prédicteur de tenue dans la durée.
4. À qui appartiennent les comptes, les numéros et les données ? La propriété du client protège la réversibilité et la valeur de l'actif constitué.

5. Quelles ressources internes le dispositif exige-t-il ? Un agent IA opéré par le fournisseur libère les équipes internes pour la conversion, plutôt que pour l'exploitation technique.

9.5 Commencez par un problème visible, puis réutilisez le même socle de connaissance

Une entreprise entre plus facilement dans l'IA agentique par un problème visible que par un programme général de transformation. Dans l'acquisition digitale, l'investissement marketing existe déjà et la perte est mesurable : formulaires sans réponse, délai de rappel, questions abandonnées, demandes non qualifiées ou contexte perdu entre canaux. Ce point d'accès fournit une preuve immédiate sans imposer aux équipes de redessiner l'entreprise entière. Une fois adopté, le même socle de connaissance, de mémoire, de droits et de mesure peut servir l'après-vente, le support, les opérations et d'autres services.

ÉTAPE	LIVRABLE DURABLE	CRITÈRE DE PASSAGE
1. Analyser	Point d'accès, coût, volumes, responsable et indicateur	Problème et seuil de succès acceptés
2. Ancrer	Sources minimales, droits, base RAG et cas de test	Pertinence et fidélité au seuil défini
3. Servir	Un canal et un parcours à valeur claire	Qualité stable sur demandes réelles
4. Prouver	Mesure métier, adoption et reprise humaine	Gain visible et décision de poursuivre
5. Étendre	Sources, outils, langues, canaux et équipes supplémentaires	Chaque extension maintient le résultat
6. Centraliser	Sacha Engine : contexte, gouvernance, traces et évaluations	Centre nerveux commun et réversible

Tableau 43 : Une feuille de route cumulative : chaque étape crée un actif réutilisable.

Le premier périmètre doit être suffisamment important pour produire de la valeur, mais assez borné pour être évalué. Une base d'avant-vente ou d'après-vente convient bien : les questions sont nombreuses, les sources existent, le transfert humain est naturel et les gains peuvent se mesurer sans confier immédiatement des décisions irréversibles à l'agent.

9.6 Mesurer le service augmenté, pas seulement la productivité

Le tableau de bord d'un système agentique RAG doit relier plusieurs niveaux. La qualité de récupération mesure si les bonnes sources sont trouvées. La qualité de réponse vérifie la fidélité et l'utilité. La qualité d'exécution contrôle les outils et l'achèvement de la tâche. La qualité de service observe ce que vivent le client et l'équipe humaine. Un indicateur isolé ne permet pas de diagnostiquer une baisse de performance.

NIVEAU	INDICATEURS RECOMMANDÉS	QUESTION DE PILOTAGE
Connaissance	Précision et rappel de recherche, fraîcheur, couverture	L'agent trouve-t-il les bonnes preuves ?
Réponse	Fidélité, citations, abstention, cohérence	Dit-il seulement ce que les sources permettent ?
Action	Succès des outils, achèvement, erreur, réversibilité	Le résultat a-t-il réellement été produit ?
Expérience	P50 et P95 de latence, satisfaction, répétition, abandon	Le service est-il fluide pour l'utilisateur ?
Humain	Qualité du transfert, temps de reprise, contexte complet	L'équipe reçoit-elle un dossier meilleur ?
Business	Conversion, résolution, chiffre d'affaires, coût par conversation	Le dispositif crée-t-il de la valeur ?

Tableau 44 : Les six niveaux à relier dans le tableau de bord d'un système agentic RAG.

9.7 Quand une réponse est mauvaise, corrigez la bonne couche au lieu de changer de modèle

La mesure devient utile lorsqu'elle permet d'identifier la couche responsable. Une réponse incorrecte peut venir d'une source absente, d'un mauvais découpage, d'une recherche imprécise, d'une règle de formulation ambiguë ou d'un outil métier indisponible. Sans cette lecture en profondeur, l'équipe risque de changer de modèle alors que le problème se trouve dans la donnée ou dans le processus.

La revue mensuelle doit donc croiser des conversations réelles, les indicateurs techniques et le retour des équipes qui reprennent les dossiers. Les corrections retenues alimentent la base de connaissances, les règles d'escalade et les jeux d'évaluation. Le système progresse ainsi par apprentissage organisationnel contrôlé, sans dépendre d'une mémoire implicite du modèle.

BOUCLE D'AMÉLIORATION

Observer les écarts, attribuer leur cause, corriger la bonne couche, rejouer les cas de test, puis mesurer à nouveau sur le trafic réel. Cette discipline transforme un pilote ponctuel en capacité d'entreprise.

10

Gouvernance, sécurité et conformité

Dès qu'un agent agit dans vos systèmes, il devient une charge de travail informatique à part entière.

Il possède une identité, des habilitations et des accès aux données. Consentement, provenance de la demande, droits d'accès, sécurité et preuve cessent donc d'être des sujets traités à part : ils fixent le niveau d'exigence du dispositif. Ce chapitre décrit les six contrôles à mettre en place, les cadres applicables aux États-Unis, dans l'Union européenne, en Israël et en Asie, puis les risques propres aux agents et leurs parades.

10.1 Les six contrôles à mettre en place

CONTRÔLE	OBJET	MISE EN ŒUVRE COURANTE
Identité de l'agent	Attribuer à l'agent une identité propre, distincte d'un compte utilisateur.	Annuaire d'entreprise, comptes de service dédiés, rotation des secrets.
Périmètre d'action	Définir de manière déterministe ce que l'agent peut faire.	Politiques d'autorisation par outil, listes d'actions permises, plafonds.
Ancrage obligatoire	Exiger une source pour toute affirmation factuelle.	Consigne d'abstention, citations systématiques, refus documenté.
Journalisation	Conserver la trace du plan, des appels d'outils, des sources et des décisions.	Observabilité de bout en bout, conservation alignée sur la politique interne.
Évaluation continue	Mesurer la qualité sur le trafic réel, à chaque évolution.	Jeux de test métier, évaluateurs de modèle, revue humaine périodique.
Escalade humaine	Transférer au bon moment, avec le contexte complet.	Seuils explicites, transfert conversationnel, reprise sans répétition.

Tableau 45 : Socle de gouvernance d'un agent en production.

Les enquêtes 2026 situent à environ 21 % la part des organisations disposant d'un modèle de gouvernance mature pour les agents autonomes. Cet écart entre déploiement et gouvernance constitue le principal facteur de risque du marché, et il explique pour partie la prévision d'annulation formulée par Gartner.

10.2 Les États-Unis : commencer par le marché, gouverner sur deux niveaux

Le marché américain ne relève pas d'un équivalent unique de l'AI Act. Au 19 août 2026, la gouvernance combine un niveau fédéral favorable à l'innovation, les règles sectorielles et de protection du consommateur déjà applicables, le cadre volontaire du NIST, puis des lois propres aux États. Pour une entreprise qui sert les États-Unis, la bonne question n'est donc pas seulement « sommes-nous conformes au niveau fédéral ? », mais « dans quels États, secteurs et décisions notre agent intervient-il ? ».

Le socle fédéral. L'Executive Order 14179 du 23 janvier 2025, puis l'Executive Order 14365 du 11 décembre 2025, placent le leadership et l'adoption de l'IA au centre de la politique nationale et recherchent un cadre national peu contraignant. Le NIST AI Risk Management Framework reste cependant la référence opérationnelle volontaire pour cartographier, mesurer, gouverner et maîtriser les risques ; son profil dédié à l'IA générative couvre notamment la provenance, les tests avant mise en service, la sécurité, la confidentialité et la gestion des incidents.

Le niveau des États. Le Texas Responsible Artificial Intelligence Governance Act est entré en vigueur le 1^{er} janvier 2026. Le Colorado a adopté en mai 2026 un cadre révisé pour les technologies de décision influençant l'emploi, le logement, le crédit, l'assurance, la santé, l'éducation ou les services essentiels ; ses obligations centrales sont prévues à partir du 1^{er} janvier 2027. D'autres États ajoutent leurs règles de transparence, de données, de biométrie ou de décisions à conséquence. L'architecture doit donc pouvoir activer des politiques, journaux, notices et seuils humains selon la juridiction, sans reconstruire l'agent.

Sources officielles : Maison-Blanche, EO 14179 et cadre national · NIST, AI RMF · Texas, HB 149 · Colorado, SB26-189.

10.3 L'Union européenne : un cadre horizontal fondé sur le risque

Une organisation établie dans l'Union européenne, ou dont le système concerne des personnes situées dans l'Union, compose avec deux corpus complémentaires. Le RGPD gouverne les données personnelles ; l'AI Act gouverne les usages, les rôles et les risques du système d'IA. Depuis le 2 août 2026, les principales dispositions générales et les obligations de transparence s'appliquent, tandis que certaines obligations relatives aux systèmes à haut risque suivent le calendrier 2027-2028 publié par la Commission.

- RGPD. Base légale, finalité, minimisation, information des personnes, durée de conservation, sécurité, sous-traitance et transferts internationaux. Un agent conversationnel collecte des données personnelles par construction : chaque canal, mémoire et connecteur doit donc être inclus dans la cartographie du traitement.
- AI Act. Transparence lorsqu'une personne interagit avec une IA, documentation proportionnée au niveau de risque, gouvernance des données, traçabilité et supervision humaine. Pour un agent commercial ou de service, la qualification du cas d'usage compte davantage que le nom du modèle : informer un prospect n'emporte pas les mêmes obligations qu'une décision influençant l'emploi, le crédit, l'assurance ou l'accès à un service essentiel.
- Souveraineté et isolation. L'hébergement régional et l'isolation des données par client répondent à une exigence devenue courante dans les appels d'offres européens, particulièrement dans les secteurs régulés.

Une bonne pratique se dégage des déploiements observés : annoncer explicitement la nature du dispositif au premier message, proposer à tout moment le passage vers un interlocuteur humain, et journaliser ces éléments. Cette transparence améliore par ailleurs la performance commerciale, l'étude Kantar relevant que 79,3 % des consommateurs recherchent une preuve de légitimité avant d'engager la conversation.

Sources officielles : RGPD, règlement (UE) 2016/679 · Commission européenne, calendrier d'application de l'AI Act.

10.4 Israël, Inde et Asie : des cadres distincts, une architecture adaptable

Les marchés asiatiques ne forment pas un bloc réglementaire homogène. Israël, l'Inde, Singapour et le Japon privilégient des cadres favorables à l'innovation, complétés par les lois sur les données, la cybersécurité et les secteurs. La Chine applique des obligations plus directement contraignantes aux services génératifs accessibles au public. Une architecture internationale doit donc séparer le socle commun — identité, droits, preuve, journalisation, sécurité et reprise humaine — des politiques activées par pays, canal et cas d'usage.

MARCHÉ	CADRE AU 19 AOÛT 2026	CONSÉQUENCE POUR L'ENTREPRISE
Israël	Approche sectorielle et favorable à l'innovation ; protection de la vie privée, cybersécurité, lignes directrices d'usage responsable ; signature de la Convention-cadre du Conseil de l'Europe sur l'IA.	Documenter la finalité, protéger les données, sécuriser l'infrastructure et maintenir le contrôle humain, sans figer l'innovation.
Inde	India AI Governance Guidelines publiées en novembre 2025 ; Digital Personal Data Protection Act et Rules 2025 pour les données personnelles.	Consentement et information adaptés, minimisation, sécurité, responsabilité et localisation analysées selon le service.
Singapour	Model AI Governance Framework ; cadre spécifique à l'IA agentique publié par l'IMDA en janvier 2026.	Limiter les pouvoirs de l'agent, tester les actions, maintenir la responsabilité humaine et préparer l'incident.
Japon	AI Guidelines for Business, version 1.2 publiée par METI et MIC en avril 2026 ; approche de gestion du risque et de responsabilité partagée.	Évaluation continue, transparence, sécurité, droits des parties et gouvernance sur tout le cycle de vie.
Chine	Mesures administratives contraignantes sur les services d'IA générative accessibles au public, règles de deep synthesis et obligations d'enregistrement pour certains algorithmes.	Vérifier le périmètre public, les contenus, les données, les fournisseurs locaux, les dépôts et l'exploitation avant lancement.

Tableau 46 : Un socle de gouvernance commun, des politiques locales activées selon le marché. Cette synthèse ne remplace pas une analyse juridique du cas d'usage.

Sources officielles : Israël, guide d'usage responsable de l'IA · Inde, DPDP Rules 2025 · Singapour, Agentic AI Framework · Japon, AI Guidelines for Business · Chine, mesures sur l'IA générative.

10.5 Les risques propres aux agents

RISQUE	DESCRIPTION	CONTRE-MESURE
Injection de consigne	Un contenu externe détourne l'agent de son objectif via une instruction dissimulée.	Traiter tout contenu récupéré comme une donnée, jamais comme une instruction ; cloisonner les outils sensibles.
Fuite de données	Un agent expose une information au-delà des habilitations du demandeur.	Filtres de sécurité au niveau de l'index, contrôle d'accès propagé jusqu'à la recherche.
Dérive de la base	La base de connaissance vieillit et l'agent cite des informations périmées.	Synchronisation planifiée, dates de validité, alertes sur documents obsolètes.
Erreur d'attribution	Dans une chaîne multi-agents, l'origine d'une erreur devient difficile à établir.	Attribution par outil et journalisation fine, recommandées par la littérature d'évaluation.
Emballement des coûts	Les boucles de raisonnement multiplient les appels de modèle.	Effort proportionné à la difficulté, plafonds par conversation, suivi du coût unitaire.

Tableau 47 : Risques spécifiques aux architectures agentiques et parades associées.

11

Les erreurs à éviter avant et pendant le déploiement

Huit anti-modèles observés, et la correction qui va avec chacun.

Klarna a dû réembaucher des humains après avoir tout confié à l'IA. 11x a été pris à gonfler ses chiffres. Sur le terrain, les IA commerciales tiennent encore moins souvent leurs promesses que les humains. Ces échecs ne veulent pas dire que l'IA agentique ne marche pas : ils montrent huit erreurs précises, corrigées une par une dans ce chapitre.

ANTI-MODÈLE	SYMPTÔME	CORRECTION
Le pilote sans indicateur	Le projet démontre une capacité technique sans engagement chiffré.	Fixer l'indicateur financier et le seuil de succès dès le cadrage.
La base de connaissance négligée	Réponses approximatives attribuées au modèle, alors que la source est en cause.	Auditer, nettoyer et attribuer une propriété éditoriale à la base.
L'autonomie totale d'emblée	Exécution agentique intégrale sans escalade, satisfaction en baisse.	Démarrer au niveau supervisé, élargir sur preuve chiffrée.
Le volume comme stratégie	Multiplication des envois, réputation d'expéditeur dégradée.	Privilégier la qualification et l'ancrage documentaire.
La sophistication uniforme	Latence et coût élevés sur des questions simples.	Calibrer l'effort de raisonnement sur la difficulté réelle.
La dépendance créée par le prestataire	Comptes, numéros et données détenus par le fournisseur.	Exiger la propriété des comptes et la portabilité des données.
L'agent isolé	Un agent par service, sans contexte partagé.	Contexte centralisé entre canaux et entre agents.
La mesure abandonnée après le lancement	Qualité en dérive silencieuse.	Évaluation continue sur trafic réel, revue mensuelle.

Tableau 48 : Huit anti-modèles observés et leur correction.

Une enquête publiée par Salesforce en février 2026 auprès de 1 050 responsables informatiques donne la mesure d'un de ces points : les organisations exploitent en moyenne 12 agents, chiffre attendu en hausse de 67 % en deux ans, et la moitié de ces agents fonctionnent en silos plutôt qu'en systèmes coordonnés. La question de l'orchestration se posera donc avant celle du nombre.

12

Là où la valeur se perd, et par où elle revient

Les modèles, les outils et les architectures ne créent pas de valeur par eux-mêmes.

C'est leur intégration dans les processus, les équipes et la gouvernance de l'entreprise qui transforme le potentiel en résultat durable. Les trois ruptures décrites en ouverture convergent ici : l'immédiateté change l'attente de service, les agents personnels déplacent l'interface entre les individus et les organisations, et le consentement, la provenance et la preuve deviennent des données opérationnelles.

L'IA générative est entrée dans les usages du grand public ; son intégration structurée dans l'entreprise commence à peine. L'enjeu n'est donc pas de distribuer un outil conversationnel supplémentaire, mais de relier les LLM à la connaissance, aux droits, aux canaux, aux outils, aux équipes et aux objectifs économiques, puis de construire cette capacité par étapes, sur le trafic réel.

1. Traiter le modèle comme un composant. La performance provient de l'architecture construite autour du modèle, ce que les travaux du MIT établissent sans ambiguïté.
2. Ancrer chaque réponse. Une affirmation sans source constitue un risque ; une affirmation citée constitue un actif.
3. Donner à l'agent la capacité d'agir. La valeur se matérialise au moment où l'agent écrit dans les systèmes métiers et boucle la tâche.
4. Laisser le raisonnement piloter la recherche. La décomposition, le routage et l'itération produisent le gain d'exactitude mesuré par la recherche académique.
5. Calibrer l'effort. Réponse directe sur le simple, itération sur le complexe : la latence et le coût se maîtrisent par la proportionnalité.
6. Investir dans la donnée avant l'agent. La préparation documentaire détermine le résultat davantage que le choix du fournisseur.
7. Assumer l'escalade humaine. Un taux de transfert stable et cadré protège la marque mieux qu'une exécution agentique intégrale.
8. Mesurer en continu. Fidélité aux sources, taux de résolution, coût unitaire et satisfaction se surveillent sur le trafic réel, pas sur un jeu de démonstration.

9. Standardiser les connexions. Un protocole ouvert et gouverné collectivement réduit le coût d'intégration et le risque d'enfermement.
10. Commencer là où la valeur se perd. La demande entrante offre un premier terrain directement mesurable : elle permet de prouver la valeur sur un périmètre réel, puis d'étendre l'architecture par étapes à mesure que les usages et la technologie progressent.

LE MOT DE LA FIN

Le LLM comprend et formule. Le RAG apporte les faits de l'entreprise. L'agentique conduit le parcours. L'humain conserve la responsabilité, le jugement et la relation. L'intégration commence par un point d'accès précis où la valeur se perd déjà ; elle démontre un résultat, s'étend progressivement, puis devient un centre nerveux commun. C'est ainsi que l'IA cesse d'être un outil isolé et devient une fonction à part entière de l'entreprise.

Salomon Bellaïche — Fondateur de SACHA OMON AI, rédacteur en chef du Livre blanc.

Synthèse décisionnelle : ce que le dirigeant doit retenir et décider

Tout ce livre revient à une seule question : quand une demande se présente, votre entreprise agit-elle, ou se contente-t-elle d'y répondre ?

Au terme de ce livre blanc, un constat s'impose : l'IA générative est devenue un usage de masse, mais son intégration structurée dans l'entreprise reste encore minoritaire. Le passage décisif ne consiste pas à distribuer un outil supplémentaire ; il consiste à relier les modèles à une connaissance gouvernée, à des règles, à des canaux, à des outils et à des équipes pour produire un résultat mesurable.

Deux capacités répondent à cet écart. L'architecture agentique permet au système de poursuivre un objectif, de préparer des étapes, d'appeler des outils autorisés et de transmettre au bon moment. Le RAG ancre la réponse dans la connaissance vérifiable et actualisable de l'entreprise. Le LLM comprend et formule ; le RAG apporte les faits ; l'agentique conduit le parcours.

CONVICTION DU FONDATEUR — PENSER GLOBAL, CONSTRUIRE PAR ÉTAPES

Je propose de penser l'architecture dans son ensemble — connaissance, droits, canaux, sécurité, outils et humain — puis de déployer marche par marche : besoin réel, première capacité, terrain, mesure, correction et extension. C'est une méthode de travail proposée dans ce Livre blanc, pas une norme universelle.

Quatorze chiffres pour cadrer la décision

INDICATEUR	VALEUR	SOURCE
Utilisateurs hebdomadaires de ChatGPT en 2026	900 M+	OpenAI, 2026
Organisations utilisant régulièrement l'IA dans au moins une fonction	88 %	McKinsey, State of AI, novembre 2025
Organisations ayant industrialisé au moins un système agentique	≈ 23 %	McKinsey, State of AI, novembre 2025
Pilotes d'IA générative sans impact mesurable sur le compte de résultat	95 %	MIT NANDA, The GenAI Divide, 2025
Applications d'entreprise embarquant un agent spécialisé fin 2026	40 % (contre moins de 5 % en 2025)	Gartner
Dépense agentique mondiale prévue en 2026	201,9 Md\$	Gartner
Marché « agentic AI » en propre, 2026	≈ 9 à 12 Md\$, TCAC supérieur à 40 %	Fortune Business Insights, Precedence Research, MarketsandMarkets
Marché RAG 2026	≈ 3,3 Md\$, TCAC ≈ 42 %	Next Move Strategy Consulting
Systèmes LLM composés en environnement d'entreprise utilisant une forme de RAG (ordre de grandeur)	≈ 60 %	Xu et al., « RAGOps: Operating and Managing Retrieval-Augmented Generation Pipelines », 2025 (arxiv.org/abs/2506.03401)
Projets agentiques susceptibles d'être annulés d'ici fin 2027	> 40 %	Gartner
Croissance des tâches exécutées sur Amazon Bedrock AgentCore en 6 mois	× 15	AWS, juin 2026
ARR Agentforce à la clôture de l'exercice 2026 de Salesforce	800 M\$ (+169 % sur un an)	Salesforce, résultats FY2026
Exportations de la haute technologie israélienne en 2025	≈ 85 Md\$, soit 58 % des exportations du pays	Israel Innovation Authority, 2026
Position d'Israël comme pôle mondial de financement technologique en 2025	4 ^e mondial, 1 ^{er} hors États-Unis ; 35 % des investissements vers le cœur de l'IA	Israel Innovation Authority, 2026

Tableau 49 : Quatorze indicateurs de référence. Détail et liens en annexe D.

Huit conclusions

1. L'écart de valeur provient de l'architecture, pas du modèle. Les travaux du MIT attribuent l'échec des pilotes à un déficit d'apprentissage contextuel et d'intégration aux processus, plutôt qu'à la seule performance des modèles.
2. L'agentique transforme la réponse en résultat. Un agent planifie, appelle des outils, écrit dans les systèmes métiers et boucle la tâche : la valeur se déplace de la génération de texte vers l'exécution du travail.

3. Le RAG s'impose comme une architecture de référence de l'IA d'entreprise. Il répond à la double exigence d'exactitude et de traçabilité et transforme le patrimoine documentaire en actif opérationnel.
4. Les fournisseurs convergent sur une architecture comparable. Les modèles restent remplaçables ; la connaissance, les règles, les canaux, les droits et l'orchestration deviennent les actifs durables de l'entreprise.
5. La demande entrante est un cas d'usage directement mesurable. En B2B comme en B2C, la vitesse de réponse, la qualification et la continuité du contexte permettent de traiter l'intention au moment où elle existe.
6. La base de connaissances est le premier actif IA durable. Structurer les sources, les droits et les dates de validité crée une couche réutilisable sur tous les canaux et dans les futurs cas d'usage internes.
7. La complémentarité humain-agent facilite l'adoption. L'agent étend la disponibilité, prépare les dossiers et transmet le contexte ; l'humain conserve le jugement, l'empathie, la négociation et les exceptions.
8. Penser global, construire marche par marche. Les outils, les usages et les méthodes de développement évoluent en même temps. Je recommande donc une vision d'architecte dès le départ, puis un premier périmètre réel, mesuré et étendu progressivement. C'est une méthode proposée ici ; d'autres trajectoires peuvent être pertinentes selon le contexte.

À PROPOS DE SACHA OMON AI

Qui publie ce livre blanc, comment l'entreprise travaille, puis le fondateur. Les annexes — panorama, glossaire, sources — commencent après, sur leur propre page.

À propos de SACHA OMON AI

De la connaissance d'entreprise à l'action multicanale : web, WhatsApp, téléphone, e-mail et SMS. Conçu et développé en Israël pour le marché national et les marchés internationaux.

SACHA OMON AI conçoit et opère un agent SDR IA qui passe à l'action dès la demande entrante : il comprend, qualifie, apporte la réponse ou le document attendu, propose la prochaine étape et transmet à l'équipe humaine lorsque c'est nécessaire. Cette page décrit ce que fait l'entreprise, comment elle travaille, et les quatre partis pris qui la distinguent des solutions du marché.

CE QUE NOUS FAISONS

Un seul agent, tous les canaux, opéré par nos équipes

L'agent SDR IA agentique multicanal agit dès la demande entrante sur le web, WhatsApp, le téléphone, l'e-mail et le SMS, avec un contexte centralisé, 24 heures sur 24 et dans plus de 40 langues. Une conversation commencée sur un canal se poursuit sur un autre sans que la personne ait à se répéter. Le socle technique est Sacha Engine, développé et opéré par l'entreprise.

SDR signifie sales development representative : dans une organisation commerciale, c'est la personne qui reçoit les demandes entrantes, vérifie qu'elles correspondent à l'offre, écarte celles qui n'ont pas lieu d'être et transmet aux commerciaux celles qui méritent un rendez-vous. C'est ce rôle-là que l'agent tient, sur tous les canaux à la fois. L'objectif de service est une première action en moins de deux secondes, à toute heure — là où le délai habituel se compte encore en heures, parfois en jours. Le chapitre 01 montre ce que coûte ce délai.

2 s	24/7	40+	1	0
Objectif de première action	Disponibilité continue, tous canaux	Langues prises en charge	Agent unique, un seul contexte partagé	Commission sur les ventes réalisées

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

PARTI PRIS N° 1 · LE CLÉ EN MAIN

Le marché vend des outils à configurer. Nous livrons un service qui fonctionne.

La plupart des solutions d’agent IA sont des plateformes. L’entreprise achète une licence, puis fait tout elle-même : décrire son offre, rassembler et trier sa connaissance, écrire les règles, brancher son CRM, tester, corriger, recommencer. Ce temps-là n’est presque jamais compté au moment de la décision. Il est toujours payé ensuite.

SACHA OMON AI a fait le choix inverse. Comprendre une activité, analyser ses demandes entrantes et traduire son raisonnement commercial dans un agent, c’est notre métier — ce n’est pas celui de notre client. Nous prenons donc le projet en charge de A à Z.

ÉTAPE DU PROJET	PLATFORME À CONFIGURER SOI-MÊME	SACHA OMON AI
Analyse de l'activité et des demandes entrantes	Le client	SACHA OMON AI
Base de connaissances : collecte, tri, validation	Le client	SACHA OMON AI
Règles, garde-fous, scénarios de transfert humain	Le client	SACHA OMON AI
Intégrations : CRM, agenda, catalogue, support	Le client	SACHA OMON AI
Tests, corrections et mise en production	Le client	SACHA OMON AI
Suivi, mesure et améliorations	Le client	SACHA OMON AI · chef de projet dédié
Propriété des comptes, des données et des actifs créés	Le client	Le client

Tableau 50 : Le client garde la propriété de tout. SACHA OMON AI porte la charge de travail.

Beaucoup de solutions demandent au client de tout faire. Nous avons choisi de tout faire pour lui. C’est plus simple, et c’est plus rapide.

PARTI PRIS N° 2 · L’OPTIMISATION TOUS LES MOIS

Un agent qui n’est pas revu chaque mois finit par se tromper

Ce marché avance plus vite que la capacité d’une entreprise à le suivre. Les modèles changent, les canaux changent, les usages changent, les règles changent. Une formation suivie en ligne peut être partiellement dépassée quelques semaines plus tard. Ce n’est ni un défaut d’organisation ni un manque de sérieux : c’est le rythme du marché.

C’est la leçon de plus de trente ans passés dans le développement web et l’acquisition digitale : une entreprise est forte quand elle est bien entourée. Nous appliquons ce principe à l’IA. L’agent n’est pas livré puis abandonné — il est mesuré, corrigé et mis à niveau tous les mois par nos équipes.

01**Mesurer**

Volumes traités, qualité des réponses, taux de transfert humain, rendez-vous et opportunités générés.

02**Analyser**

Questions sans réponse satisfaisante, objections nouvelles, sources manquantes ou périmées.

03**Corriger**

Mise à jour de la base de connaissances, des règles, des scénarios et des intégrations.

04**Mettre à niveau**

Montées de version, nouveaux modèles, nouveaux canaux et nouvelles capacités, sans intervention du client.

Un chef de projet dédié suit le compte. Le client n'a ni modèle à surveiller, ni version à installer, ni prestataire technique à coordonner.

PARTI PRIS N° 3 • LE MONO-LOCATAIRE, OU SINGLE TENANT**Une infrastructure dédiée par client, jamais partagée**

Deux architectures coexistent dans le logiciel d'entreprise. En multi-locataire — multi-tenant —, tous les clients d'un éditeur partagent la même application et la même base de données ; leurs données sont séparées par des règles logicielles. En mono-locataire — single tenant —, chaque client dispose de sa propre instance et de sa propre base. SACHA OMON AI a retenu le mono-locataire.

POINT DE COMPARAISON	MULTI-LOCATAIRE (MULTI-TENANT)	MONO-LOCATAIRE (SINGLE TENANT) • NOTRE CHOIX
Base de données	Partagée entre les clients de l'éditeur	Base PostgreSQL privée, propre au client
Isolation des données	Logique, gérée par le logiciel	Physique, par instance dédiée
Voisinage	L'activité d'un autre client peut peser sur la vôtre	Aucun voisinage
Configuration	Commune, options limitées	Adaptée aux règles et aux processus du client
Réversibilité	Extraction dépendante de l'éditeur	Périmètre identifiable, exportable
Conformité	À démontrer au niveau de la plateforme	À démontrer sur un périmètre délimité

Tableau 51 : Le mono-locataire coûte plus cher à opérer. Il évite de confondre l'usage d'un service avec la propriété de l'actif IA de l'entreprise.

Ce choix a une conséquence directe pour le dirigeant : la connaissance de l'entreprise, ses conversations, ses règles et les actifs constitués au cours du projet restent dans un périmètre qui lui appartient. La couche d'intelligence — le modèle — reste externe et sélectionnable, et n'accède qu'aux éléments nécessaires à la tâche autorisée. Le traitement est aligné sur le RGPD.

PARTI PRIS N° 4 • AGENTIQUE + RAG

Le RAG apporte la justesse. L’agentique apporte l’action. Il faut les deux.

Ces deux mots reviennent dans tout le livre blanc. Voici, en trois lignes, ce qu’ils recouvrent et pourquoi leur combinaison est le cœur de notre architecture.

BRIQUE	CE QU’ELLE FAIT	CE QUI MANQUE SANS ELLE
LLM seul	Comprend la langue et formule une réponse fluide	Il ne connaît ni vos offres, ni vos prix, ni vos règles : il peut inventer
RAG	Va chercher la réponse dans les documents validés de l’entreprise avant de répondre	Sans RAG, la réponse est plausible mais pas vérifiable
Agentique	Reçoit un objectif, planifie, utilise les outils, observe le résultat et corrige jusqu’au bout	Sans agentique, le système répond mais n’agit pas : pas de rendez-vous, pas de document, pas de relance

Tableau 52 : Trois briques, trois rôles distincts. Aucune ne remplace les deux autres.

POURQUOI LA COMBINAISON CHANGE LE RÉSULTAT

Le RAG donne à l’agent une connaissance juste, datée et attribuable : il répond à partir des données de l’entreprise, pas de sa mémoire d’entraînement.

L’agentique donne à cette connaissance une capacité d’action : qualifier, proposer un créneau, envoyer le document, relancer, transmettre à un commercial avec tout le contexte.

Ensemble, elles produisent ce qu’une entreprise attend réellement : une demande accompagnée en actions, du premier message jusqu’à sa conclusion — et traçable à chaque étape.

Automatisation et agentique sont souvent confondues, alors qu’elles répondent à des besoins différents. Un système d’automatisation exécute une séquence programmée à l’avance, identique à chaque déclenchement — utile pour un processus stable et répétitif. Un système agentique choisit son chemin à chaque interaction, en fonction de l’objectif et de ce qu’il observe — nécessaire dès que la demande varie, ce qui est le cas de la quasi-totalité des demandes commerciales entrantes. La même distinction vaut pour les chatbots à règles, qui répondent dans un scénario fermé prévu à l’avance (voir Tableau 20, page 49). Sacha Omon relève de la troisième catégorie : agentique, et non automatisation ou chatbot.

LE FONDATEUR

Salomon Bellaïche — trente ans d’acquisition digitale, aujourd’hui traduits dans l’IA agentique

Avant le digital, Salomon Bellaïche a passé huit ans dans la production audiovisuelle, la presse et la musique. En 1996, il crée l’une des premières agences digitales françaises consacrées à l’acquisition digitale et à la génération de demande. Depuis, il a accompagné plus de 3 000 entreprises, de la TPE aux grands comptes, dont Microsoft et L’Oréal.

Le fil conducteur est resté le même : transformer un investissement marketing en demande réelle, comprendre l’intention, qualifier le besoin et conduire la relation jusqu’au résultat commercial. Plus de trois ans de terrain dans des centres d’appels en Israël ont complété cette expérience par l’observation

directe des conversations, des objections, des transferts et des moments où l'humain doit reprendre la main.

Aujourd'hui, cette expérience est mise au service de SACHA OMON AI. Son développement représente près de quatre années de travail et l'entreprise est autofinancée depuis sa création. Sacha Engine associe LLM, RAG et orchestration agentique pour relier la connaissance de l'entreprise, ses outils, ses canaux et ses équipes. SACHA OMON AI relève d'une troisième catégorie : ni automatisation, ni chatbot.

L'automatisation exécute un chemin prévu ; le chatbot converse ; l'agentique comprend une situation, recherche la connaissance nécessaire, choisit parmi les actions autorisées et agit dans un cadre contrôlé. Conçue et développée en Israël, la solution vise le marché israélien et les marchés internationaux, avec une architecture pensée pour les États-Unis, l'Europe et l'Asie. L'objectif : comprendre chaque besoin exprimé par un prospect, un client ou un citoyen, lui apporter la bonne réponse et déclencher l'action adaptée pour le satisfaire, 24 heures sur 24 et 7 jours sur 7, grâce à l'IA.

Profil du fondateur : [linkedin.com/in/salomon-sylvain-bellaïche-8187796](https://www.linkedin.com/in/salomon-sylvain-bellaïche-8187796) · Site : www.sacha-omon.ai · Démonstration : www.sacha-omon.com

POUR ALLER PLUS LOIN

Étude stratégique : analyse de vos demandes entrantes, de vos canaux et de votre délai de première réponse, avec une estimation des opportunités aujourd'hui perdues, retardées ou insuffisamment traitées. → www.sacha-omon.ai

Cahier des charges autonome : utilisez le script des pages 88 à 92 dans votre propre IA pour cadrer votre premier cas d'usage et produire une première version du document. Aucun simulateur ni agent de cadrage en ligne n'est nécessaire.

Démonstration de 30 minutes avec étude de potentiel, sans engagement. → www.sacha-omon.com

Contact direct : sacha@sacha-omon.com

LinkedIn du fondateur : [linkedin.com/in/salomon-sylvain-bellaïche-8187796](https://www.linkedin.com/in/salomon-sylvain-bellaïche-8187796)

© 2026 SACHA OMON AI. Tous droits réservés. Citation autorisée avec mention de la source ; reproduction soumise à accord écrit. Les marques citées appartiennent à leurs détenteurs respectifs. Les données de marché sont reproduites à titre informatif et renvoient à leurs éditeurs d'origine ; elles n'ont pas fait l'objet d'une vérification indépendante par SACHA OMON AI et n'engagent que leurs auteurs respectifs. Ce document est fourni à titre purement informatif : il ne constitue ni un conseil juridique, financier ou d'investissement, ni un engagement contractuel de la part de SACHA OMON AI.

ANNEXES DU LIVRE BLANC

Panorama des éditeurs, glossaire, méthodologie, sources et repères chronologiques — pour vérifier, aller plus loin et citer ce livre blanc avec exactitude.

Annexe A — Panorama des éditeurs de modèles et de leurs moteurs

Ce livre blanc est volontairement agnostique : il décrit des architectures, pas des produits. Ce tableau recense les principaux éditeurs de modèles de langage cités dans l'ouvrage, avec leur pays d'origine et l'adresse de leur documentation officielle, afin que chaque lecteur puisse vérifier lui-même l'état des offres plutôt que s'en remettre à ce document.

ÉDITEUR	ORIGINE	MOTEURS ET FAMILLES DE MODÈLES	DOCUMENTATION OFFICIELLE
OpenAI	États-Unis	GPT, ChatGPT	openai.com/index · platform.openai.com/docs
Anthropic	États-Unis	Claude : Opus, Sonnet, Haiku	anthropic.com/news · docs.anthropic.com
Google et Google DeepMind	États-Unis	Gemini : Pro, Flash	blog.google · ai.google.dev
Microsoft	États-Unis	Copilot, Phi, Azure AI Foundry	microsoft.com/copilot · learn.microsoft.com
Meta	États-Unis	Llama, Meta AI	ai.meta.com/llama
xAI	États-Unis	Grok	x.ai/news · docs.x.ai
Amazon Web Services	États-Unis	Nova, plateforme Bedrock	aws.amazon.com/bedrock
Perplexity	États-Unis	Moteur de recherche génératif	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	Chine	DeepSeek V, DeepSeek R	api-docs.deepseek.com
Alibaba Cloud	Chine	Qwen	qwen.ai · alibabacloud.com
Moonshot AI	Chine	Kimi	moonshot.ai · platform.moonshot.ai
Zhipu AI	Chine	GLM	z.ai · bigmodel.cn
ByteDance	Chine	Doubao	volcengine.com
EPFL, ETH Zurich et le CSCS	Suisse	Apertus	apertus.ai · huggingface.co/swiss-ai

Tableau 53 : Panorama non exhaustif, arrêté au 19 août 2026. Les versions et les tarifs changent en permanence : seule la documentation officielle de chaque éditeur fait foi.

Aucun de ces éditeurs n'a participé à la rédaction de ce livre blanc, ne l'a relu ni ne l'a financé. Les architectures décrites ici — agentique et RAG — sont indépendantes du fournisseur retenu : c'est précisément l'intérêt d'une architecture multi-modèles, décrite au chapitre 7.

Annexe B — Glossaire : termes et définitions

TERME	DÉFINITION
A2A (Agent2Agent)	Protocole ouvert publié par Google en avril 2025, gouverné depuis par la Linux Foundation, permettant à des agents IA de se découvrir et de coordonner des actions entre eux.
ACP (Agentic Commerce Protocol)	Standard ouvert codéveloppé par OpenAI et Stripe, qui structure les échanges entre un agent IA, un acheteur et un marchand pour conclure un achat.
Agent agentique	Système qui reçoit un objectif, planifie, appelle des outils, observe, corrige et conduit la tâche jusqu'à un état terminal.
Agentic RAG	Architecture combinant raisonnement agentique et recherche documentaire : l'agent décide quoi chercher, où, combien de fois, et évalue la suffisance des preuves.
AP2 (Agent Payments Protocol)	Extension d'A2A publiée par Google, qui représente chaque paiement exécuté par un agent sous la forme de mandats signés cryptographiquement, prouvant l'autorisation de l'utilisateur.
Apertus (EPFL, ETH Zurich, CSCS)	Famille de modèles à poids ouverts et multilingue, développée par un consortium académique suisse (EPFL, ETH Zurich et le Centre national suisse de calcul scientifique). Documentation : apertus.ai et huggingface.co/swiss-ai .
Appel d'outil (tool call)	Invocation structurée d'une fonction externe par l'agent : rechercher un document, lire un calendrier ou créer une opportunité CRM.
Architecture multi-modèles	Architecture capable de mobiliser plusieurs fournisseurs ou familles de modèles sans enfermer la connaissance et les processus de l'entreprise dans un seul moteur.
Base de connaissances	Ensemble gouverné de sources validées, attribuées, datées et accessibles selon des droits explicites, utilisé par un ou plusieurs systèmes d'IA.
Canal conversationnel	Interface par laquelle l'utilisateur échange avec l'agent : web, WhatsApp, téléphone, e-mail, SMS ou autre messagerie.
ChatGPT et GPT (OpenAI)	ChatGPT est l'application grand public d'OpenAI (États-Unis) ; GPT désigne la famille de modèles qui la fait fonctionner. Documentation : openai.com/index et platform.openai.com/docs .
Chunking (découpage)	Segmentation des documents en fragments cohérents avant vectorisation.

TERME	DÉFINITION
Claude (Anthropic)	Famille de modèles de langage développée par Anthropic (États-Unis), déclinée en Opus, Sonnet et Haiku. Documentation : anthropic.com/news et docs.anthropic.com .
Copilot (Microsoft)	Gamme d'assistants IA intégrés aux produits Microsoft (États-Unis), adossée à la plateforme Azure AI Foundry. Documentation : microsoft.com/copilot et learn.microsoft.com/azure/ai-foundry .
Couche d'intelligence (LLM)	Composant chargé de comprendre, raisonner, planifier, synthétiser et formuler ; il ne remplace ni la base de connaissances, ni la mémoire, ni les autorisations.
DeepSeek	Éditeur chinois de modèles à poids ouverts, connu pour les familles V et R et pour son rapport performance-coût. Documentation : api-docs.deepseek.com .
Doubao (ByteDance)	Famille de modèles et assistant grand public développés par ByteDance (Chine), très diffusés sur le marché intérieur chinois. Documentation : volcengine.com .
Embedding	Représentation numérique d'un texte permettant la comparaison sémantique.
Escalade (transfert humain)	Passage contrôlé de l'agent à une personne, accompagné de l'identité, de l'historique, des sources consultées et de la prochaine action recommandée.
Fidélité (faithfulness)	Mesure de l'adéquation entre la réponse produite et les sources fournies.
Garde-fou	Contrôle déterministe encadrant le comportement de l'agent, indépendamment du modèle.
Gemini (Google)	Famille de modèles multimodaux de Google et Google DeepMind (États-Unis), déclinée en versions Pro et Flash. Documentation : blog.google et ai.google.dev .
GLM (Zhipu AI)	Famille de modèles à poids ouverts développée par Zhipu AI (Chine). Documentation : bigmodel.cn et z.ai .
GraphRAG	Variante du RAG exploitant un graphe de connaissance pour relier les entités entre elles.
Grok (xAI)	Famille de modèles développée par xAI (États-Unis), intégrée à la plateforme X et accessible par API. Documentation : x.ai/news et docs.x.ai .
Hallucination	Production d'une information plausible et fausse, en l'absence de source.
Kimi (Moonshot AI)	Famille de modèles à long contexte développée par Moonshot AI (Chine). Documentation : moonshot.ai et platform.moonshot.ai .
Llama (Meta)	Famille de modèles à poids ouverts publiée par Meta (États-Unis), largement utilisée pour les déploiements auto-hébergés. Documentation : ai.meta.com/llama .
LLM	Modèle de langage de grande taille, moteur de compréhension et de génération de texte.

TERME	DÉFINITION
MCP (Model Context Protocol)	Standard ouvert de connexion entre systèmes d'IA et outils, gouverné depuis décembre 2025 par l'Agentic AI Foundation.
Mémoire persistante	Conservation du contexte et de l'historique au-delà d'une session.
Mistral AI	Éditeur français de modèles de langage, à poids ouverts et propriétaires, avec l'assistant Le Chat et une API d'agents. Principal acteur européen des modèles de fondation. Documentation : mistral.ai/news et docs.mistral.ai .
Mono-locataire	Architecture dans laquelle chaque client dispose d'une infrastructure et d'une base dédiées.
Multi-sauts (multi-hop)	Raisonnement exigeant plusieurs recherches successives, chacune informée par la précédente.
Nova et Bedrock (Amazon)	Nova est la famille de modèles d'Amazon (États-Unis) ; Bedrock est la plateforme AWS qui donne accès à plusieurs familles de modèles. Documentation : aws.amazon.com/bedrock .
Opt-in (consentement préalable)	Régime dans lequel une sollicitation commerciale n'est licite qu'après accord explicite et préalable de la personne démarchée, par opposition à l'opt-out, où elle est licite sauf refus exprimé à l'avance.
Orchestration	Couche qui coordonne le raisonnement, les règles, la mémoire, la recherche documentaire, les outils et l'arrêt du processus.
Perplexity	Moteur de recherche fondé sur l'IA générative (États-Unis), qui produit des réponses accompagnées de leurs citations. Documentation : perplexity.ai .
Première réponse utile	Premier retour qui fait réellement progresser la demande ; elle se distingue d'un accusé de réception sans valeur opérationnelle.
Qwen (Alibaba Cloud)	Famille de modèles à poids ouverts développée par Alibaba Cloud (Chine), parmi les plus reprises au monde pour l'auto-hébergement. Documentation : qwen.ai et alibabacloud.com .
RAG	Retrieval-Augmented Generation : génération de réponses ancrées dans une base documentaire interrogée au moment de la question.
RAG dédié	Base de connaissances et pipeline de recherche isolés pour une entreprise ou une entité, sans mélange de données avec d'autres clients.
Reranking	Reclassement fin des passages candidats par un modèle spécialisé, après la recherche initiale.
Routage de modèles	Sélection par l'orchestrateur du modèle le plus adapté à une tâche selon la langue, la latence, la profondeur de raisonnement, le coût et les contraintes de données.

TERME	DÉFINITION
Source d'autorité	Document, système ou propriétaire éditorial reconnu comme référence lorsqu'il existe plusieurs versions d'une information.
Speed-to-lead	Délai entre l'arrivée d'une demande entrante et la première réponse utile.
Supervision humaine	Dispositif de validation ou d'intervention humaine intégré au parcours de l'agent.
Token	Unité élémentaire utilisée par un LLM pour traiter l'information. Les API facturent généralement les tokens d'entrée et de sortie ; le volume dépend du contexte transmis, des étapes de raisonnement et de la réponse générée. Le token est une unité importante du coût du LLM, mais pas l'unique coût d'un système agentique.
Unité de travail agentique	Tâche discrète accomplie par un agent : décision prise, enregistrement mis à jour, processus déclenché.
Web agentique	Évolution émergente d'Internet dans laquelle des agents IA se découvrent, s'authentifient et échangent directement entre eux, au-delà de la simple navigation humaine sur des sites web.

Tableau 54 : Glossaire des termes employés dans ce livre blanc, en une seule séquence alphabétique.

Annexe C — Méthodologie

Ce livre blanc repose exclusivement sur des sources publiques, consultées et vérifiées jusqu'au 19 août 2026. Quatre familles de sources ont été mobilisées.

1. Cabinets d'analyse : Gartner, McKinsey, IDC, Deloitte, Forrester et cabinets d'études de marché spécialisés, pour les données d'adoption, de dépense et de prévision.
2. Communications officielles des éditeurs : annonces produit, communications de résultats et pages d'études de cas d'Amazon, AWS, Salesforce, Microsoft, Google, OpenAI et Anthropic.
3. Recherche académique : publications arXiv et revues à comité de lecture portant sur le RAG, l'agentique RAG et l'évaluation de la fidélité, retenues pour leurs résultats chiffrés et reproductibles.
4. Presse économique et technologique de référence : TechCrunch, CNBC, Axios, Fortune, Forbes, GeekWire, pour les faits d'entreprise datés.

Trois principes de traitement ont été appliqués. Les chiffres issus des éditeurs sont signalés comme tels, la communication commerciale appelant une lecture prudente. Les écarts entre cabinets sont exposés plutôt qu'arbitrés, la divergence portant sur les périmètres mesurés. Les données invérifiables ou attribuées à des organismes sans existence documentée ont été écartées, y compris lorsqu'elles circulent largement.

Annexe D — Sources et articles cliquables

Cabinets d'analyse, recherche et presse

ORGANISATION	RÉFÉRENCE	ACCÈS
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 (juil. 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	Prévisions dépense agentique 2026, adoption applicative, taux d'annulation 2027	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
Auteurs multiples	Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG	arxiv.org/abs/2501.09136
Auteurs multiples	An Empirical Study of Multi-Agent RAG (MARAU)	arxiv.org/abs/2507.11272
Auteurs multiples	Agentic Hybrid RAG for Evidence-Grounded Analysis	arxiv.org/abs/2606.10381
Auteurs multiples	Agent-Orchestrated Adaptive RAG: A Comparative Study	arxiv.org/abs/2606.05658

ORGANISATION	RÉFÉRENCE	ACCÈS
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	Estimations du marché agentic AI	fortunebusinessinsights.com/agentic-ai-market-114233 (Fortune Business Insights) ; precedenceresearch.com/agentic-ai-market (Precedence Research) ; Deloitte : aucune page de rapport dédiée identifiée publiquement
The Business Research Company	AI Sales Development Representative (SDR) Global Market Report (juil. 2026)	thebusinessresearchcompany.com/report/artificial-intelligence-ai-sales-development-representative-sdr-global-market-report
Kantar pour Meta	The State of Business Messaging (11 056 répondants, 22 marchés)	whatsappbusiness.com/resources/resource-library/state-of-business-messaging/
Sacra	Analyses de revenus Sierra et Decagon (2026)	sacra.com/research/sierra-vs-decagon/
Juniper Research	Trafic mondial de messagerie professionnelle 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 (avr. 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 (févr. 2026)	researchandmarkets.com/reports/6217876/united-states-digital-ad-spend-market-size-and
HubSpot	Part des sollicitations téléphoniques reçues hors horaires standards	hubspot.com
Foureyes / Ringlead	Internet Lead Response Statistics for Dealerships : étude sur 22 500 concessionnaires, 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 (avr. 2025)	reuters.com/business/media-telecom/verizon-says-google-ai-customer-service-agents-has-led-sales-jump-2025-04-09/

ORGANISATION	RÉFÉRENCE	ACCÈS
Reuters	Teleperformance shares plunge on AI disruption concerns (28 févr. 2024)	reuters.com/technology/teleperformance-shares-plunge-ai-disruption-concerns-2024-02-28/
Euronext Paris	TP SE : cours et historique de marché, clôture du 7 août 2026	live.euronext.com/en/product/equities/FR0000051807-XPAR
TP	Informations investisseurs, résultats S1 2026 et positionnement TP.AI	tp.com/media/1yxhdsni/tp-press-release-h1-2026-results.pdf
Financial Times	Hedge funds bet against call-centre groups as AI disruption accelerates (juin 2026)	ft.com
TechCrunch / CNBC / Axios	Prometheus : levée de 12 Md\$ à 41 Md\$ de valorisation (juin 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, épisode 149, 29 août 2025	podcasts.apple.com/us/podcast/ep-149-marc-benioff-ceo-salesforce-predicts-half-of/id1606770839?i=1000724017332
Salesforce / SEC	Form 10-Q, trimestre clos le 30 avril 2026 — acquisition de Qualified.com, Inc. ; juste valeur de la contrepartie : 1,2 Md\$, principalement 1,1 Md\$ en numéraire	www.sec.gov/Archives/edgar/data/1108524/000110852426000127/0001108524-26-000127-index.htm
Digital Applied	AI SDR Real Performance: 100K Email Analysis 2026 (26 avr. 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 (15 août 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

Tableau 55 : Sources d'analyse, de recherche et de presse.

Éditeurs, plateformes et documentation officielle

ORGANISATION	RÉFÉRENCE	ACCÈS
Salesforce	Agentforce Metrics ; résultats FY2026 ; étude de cas site d'aide	salesforce.com/agentforce/metrics
Salesforce	Agentforce launch at Dreamforce (17 sept. 2024)	salesforce.com/news/stories/dreamforce-24-recap/

ORGANISATION	RÉFÉRENCE	ACCÈS
Salesforce	Agentic Enterprise Index H1 2025 (17 sept. 2025) et 2025-2026 (7 août 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 (17 déc. 2025)	salesforce.com/news/stories/salesforce-signs-definitive-agreement-to-acquire-qualified/
Salesforce	State of Sales 2026, 7 ^e édition (févr. 2026)	salesforce.com/sales/state-of-sales
Qualified	Études de cas Piper, dont Demandbase	qualified.com/customers
IBM Think	Étude sur la redéfinition de la vente par les agents SDR IA (avr. 2026)	ibm.com/think/topics/ai-sdr
Amazon / AWS	AWS Summit New York 2026 : agents, Knowledge Bases managées, AWS Context, Web Search	aboutamazon.com/news/aws/aws-summit-nyc-2026-ai-agents
Amazon	Résultats Q2 2026 : commentaire d'Andy Jassy sur 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 (13 sept. 2023) et disponibilité (28 nov. 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 (16 juil. 2025), disponibilité (13 oct. 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 (3 mars 2026), Evaluations (31 mars 2026), Policy with Guardrails (17 juin 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/

ORGANISATION	RÉFÉRENCE	ACCÈS
AWS	Agentic RAG in Amazon Q Business (14 août 2025) ; Hybrid RAG (6 avr. 2026) ; Managed Knowledge Base (16 juil. 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 (30 juil. 2026)	docs.aws.amazon.com/bedrock
Anthropic	Introducing Claude (14 mars 2023) ; Claude 3 (4 mars 2024) ; Claude 3.5 Sonnet (21 juin 2024)	anthropic.com/news/introducing-claude ; anthropic.com/news/claude-3-family ; anthropic.com/news/claude-3-5-sonnet
Anthropic	Contextual Retrieval (19 sept. 2024) ; Building Effective Agents (19 déc. 2024)	anthropic.com/engineering/contextual-retrieval ; anthropic.com/engineering/building-effective-agents
Anthropic	Introducing the Model Context Protocol (25 nov. 2024) ; MCP donation and Agentic AI Foundation (9 déc. 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 (24 févr. 2025) ; Claude 4 (22 mai 2025) ; multi-agent research system (13 juin 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 (29 sept. 2025) ; Anthropic Labs (13 janv. 2026)	anthropic.com/engineering/effective-context-engineering-for-ai-agents ; anthropic.com/news/introducing-anthropic-labs
Anthropic	Claude Opus 4.6 (5 févr. 2026) ; Sonnet 4.6 (17 févr. 2026) ; Fable 5 et Mythos 5 (9 juin 2026) ; Sonnet 5 (30 juin 2026) ; Opus 5 (24 juil. 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 (11 juin 2018) ; GPT-2 (14 févr. 2019) ; GPT-3 (2020) ; OpenAI API (11 juin 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/

ORGANISATION	RÉFÉRENCE	ACCÈS
OpenAI	WebGPT (16 déc. 2021) ; InstructGPT (27 janv. 2022) ; ChatGPT (30 nov. 2022) ; GPT-4 (14 mars 2023)	openai.com/index/webgpt/ ; openai.com/index/instruction-following/ ; openai.com/index/chatgpt/ ; openai.com/index/gpt-4-research/
OpenAI	Function calling (13 juin 2023) ; GPT-4o (13 mai 2024) ; o1-preview (12 sept. 2024) ; o3 et o4-mini (16 avr. 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 (11 mars 2025) ; gpt-oss (5 août 2025) ; GPT-5 (7 août 2025) ; AgentKit (6 oct. 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 (11 mars 2026) ; Agents SDK (15 avr. 2026) ; Presence (22 juil. 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 (5 mars 2026) ; GPT-5.5 (23 avr. 2026) ; GPT-5.6 Sol preview (26 juin 2026) ; GPT-5.6 (9 juil. 2026) ; Sol et Luna dans ChatGPT (6 août 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 – févr. 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 (6 déc. 2023) ; Gemini 2.0 (11 déc. 2024) ; Gemini 2.5 (25 mars 2025) ; Gemini 3 (18 nov. 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/

ORGANISATION	RÉFÉRENCE	ACCÈS
Google	Gemini 3.1 Pro (19 févr. 2026) ; Gemini 3.5 Flash (19 mai 2026) ; Gemini 3.6 Flash (21 juil. 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 (9 avr. 2025) ; Agent Payments Protocol AP2 (16 sept. 2025) ; Agentic Resource Discovery (17 juin 2026)	developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/ ; 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 (30 juil. 2025) ; Gemini Embedding 2 (mars et 30 avr. 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 janv. 2025) ; Gemini Enterprise Agent Platform (22 avr. 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 (19 mai 2026) ; contrôles de budget et audits d'outils (28 juil. 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 et partenaires	Agent Plugins 1.0.0 (6 août 2026)	developers.googleblog.com/agent-plugins-package-your-skills-tools-and-more/
Google	Annonces Google I/O 2026 : agent d'achat personnel, Business Agent (mai 2026)	blog.google/innovation-and-ai/technology/ai
Meta AI	Dense Passage Retrieval (2020) ; LLaMA (24 févr. 2023) ; Llama 2 (18 juil. 2023) ; Llama 3 (18 avr. 2024) ; Llama 3.1 (23 juil. 2024) ; Llama 4 (5 avr. 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/

ORGANISATION	RÉFÉRENCE	ACCÈS
Microsoft Research	AutoGen (25 sept. 2023) ; GraphRAG (avr. 2024) ; GraphRAG 1.0 (16 déc. 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 (27 sept. 2023) ; Agents API (27 mai 2025) ; Mistral 3 (2 déc. 2025) ; Medium 3.5 et Work mode (22 mai 2026)	mistral.ai/news/announcing-mistral-7b/ ; mistral.ai/news/agents-api
xAI	Grok-2 (13 août 2024) ; Grok 3 (19 févr. 2025) ; Grok 4 (9 juil. 2025) ; Grok Models and Agent Tools	x.ai/news , docs.x.ai
DeepSeek	DeepSeek-R1 (20 janv. 2025) ; V3.2 (1 ^{er} déc. 2025) ; V4 Preview (24 avr. 2026)	api-docs.deepseek.com/news/news250120
Model Context Protocol	Specification 2026-07-28 (28 juil. 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 janv. 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

Tableau 56 : Sources éditeurs, plateformes et documentation officielle.

Sources juridiques et réglementaires

JURIDICTION	RÉFÉRENCE	ACCÈS
France	Loi n° 2025-594 du 30 juin 2025 ; décret n° 2026-662 du 23 juillet 2026 ; article L223-1 du Code de la consommation	legifrance.gouv.fr/jorf/id/JORFTEXT000051824277 ; legifrance.gouv.fr/jorf/id/JORFTEXT00054483939 ; legifrance.gouv.fr/codes/article_lc/LEGIARTI000051830285/2026-08-11
France	DGCCRF : les règles du démarchage téléphonique s'imposant aux entreprises	economie.gouv.fr/dgccrf

JURIDICTION	RÉFÉRENCE	ACCÈS
France	CNIL : prospection commerciale par courrier électronique, SMS, MMS et automate d'appel	cnil.fr/fr/la-prospection-commerciale-par-courrier-electronique-sms-mms-et-automate-dappel
Union européenne	RGPD, règlement (UE) 2016/679 ; AI Act : calendrier et FAQ d'application	eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32016R0679 ; digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act
Commission européenne	Retrait de la proposition de règlement ePrivacy (16 juillet 2025)	Journal officiel de l'UE, C/2025/5423
États-Unis	FCC : AI-generated voices in robocalls (8 févr. 2024)	fcc.gov/document/fcc-confirms-tcpa-applies-ai-technologies-generate-human-voices
États-Unis	Insurance Marketing Coalition v. FCC, 11 ^e circuit fédéral, 24 janvier 2025	décision judiciaire publique
États-Unis	Executive Orders 14179 et 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
Israël	Guide d'usage responsable de l'IA ; Convention-cadre du Conseil de l'Europe	harhayed.gov.il/data-and-ai/policy/responsible-use-new/ ; gov.il
Singapour / Japon	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
Chine	Mesures administratives provisoires sur les services d'IA générative	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/

Tableau 57 : Sources juridiques et réglementaires.

Annexe E — Sources cliquables de l'ouverture stratégique

Les faits et les chiffres cités dans l'ouverture stratégique s'appuient sur des sources primaires et officielles ; toute projection ou interprétation relève de la ligne éditoriale de Salomon Bellaïche. Prometheus est contre-vérifié par deux médias de référence, faute de publication institutionnelle équivalente disponible.

[1] Zendesk CX Trends 2026 • 2026 (zendesk.com/blog/zendesk-insights/trends/cx-trends-2026-contextual-intelligence/)

Signal sur l'accélération des attentes de réponse, la disponibilité permanente et la continuité multimodale.

[2] Google I/O 2026 — Welcome to the agentic Gemini era • 19 mai 2026 (blog.google/innovation-and-ai/sundar-pichai-io-2026/)

Google présente Gemini Spark comme un agent personnel qui agit au nom de l'utilisateur, sous sa direction.

[3] Google — The Gemini app becomes more agentic • 19 mai 2026 (blog.google/innovation-and-ai/products/gemini-app/next-evolution-gemini-app/)

Gemini Spark est décrit comme un agent personnel permanent qui travaille en arrière-plan et utilise les outils Google.

[4] Meta — The Future is for Everyone • 10 août 2026 (about.fb.com/news/2026/08/the-future-is-for-everyone/)

Mark Zuckerberg expose une vision de superintelligence personnelle mise entre les mains des individus.

[5] Meta — Meta Business Agent • 3 juin 2026 (about.fb.com/news/2026/06/meta-business-agent/)

Meta annonce plus d'un million d'entreprises utilisant Business Agent et plus d'un milliard de fils actifs quotidiens avec des entreprises sur WhatsApp, Messenger et Instagram.

[6] Meta — WhatsApp message translations • 23 sept. 2025 (about.fb.com/news/2025/09/introducing-message-translations-whatsapp/)

Meta indique plus de 3 milliards d'utilisateurs de WhatsApp dans plus de 180 pays.

[7] Microsoft — Human-led, agent-operated • 16 avr. 2026 (microsoft.com/insidettrack/blog/becoming-a-frontier-firm-a-guide-for-deploying-ai-agents-based-on-our-experience-at-microsoft/)

Microsoft décrit un modèle de travail où l'IA opère mais où l'humain dirige.

[8] Microsoft 365 Copilot, human agency • 5 mai 2026 (microsoft.com/en-us/microsoft-365/blog/2026/05/05/microsoft-365-copilot-human-agency-and-the-opportunity-for-every-organization/)

Microsoft formule le cadre « human-led, agent-operated, outcome-driven ».

[9] OpenAI — ChatGPT Work • 9 juil. 2026 (openai.com/index/chatgpt-for-your-most-ambitious-work/)

ChatGPT Work rassemble informations, applications et fichiers pour mener des tâches longues et multi-étapes.

[10] OpenAI — Workspace agents • 22 avr. 2026 (openai.com/index/introducing-workspace-agents-in-chatgpt/)

OpenAI présente des agents de travail qui exécutent des processus complexes dans le cloud.

[11] OpenAI — Operator • 23 janv. 2025 (openai.com/index/introducing-operator/)

Operator est présenté comme un agent capable d'utiliser le web pour exécuter une tâche de manière indépendante.

[12] xAI — Grok Voice Agent Builder • juil. 2026 (x.ai/news/grok-voice-agent-builder)

xAI propose un constructeur sans code d'agents vocaux de production.

[13] xAI — Voice API • 2026 (x.ai/api/voice)

La Voice API de Grok prend en charge conversation en temps réel, outils et recherche.

[14] xAI — Connect Grok to Hermes Agent • 15 mai 2026 (x.ai/news/grok-hermes)

Grok peut être utilisé dans Hermes Agent, l'agent open source et persistant de Nous Research, qui crée une mémoire longue entre les sessions et peut se connecter à WhatsApp, Discord, Telegram, Signal et d'autres services de messagerie.

[15] xAI — Agent Tools API • 19 nov. 2025 (x.ai/news/grok-4-1-fast)

xAI décrit des agents de production spécialisés dans l'usage d'outils et la recherche agentique.

[16] Apple — Siri AI • 8 juin 2026 (apple.com/newsroom/2026/06/apple-introduces-siri-ai-a-profoundly-more-capable-and-personal-assistant/)

Siri AI exploite le contexte personnel dans Messages, e-mails, Photos et d'autres applications, et peut agir dans les applications.

[17] Amazon — Andy Jassy on generative AI • 17 juin 2025 (aboutamazon.com/news/company-news/amazon-ceo-andy-jassy-on-generative-ai)

Andy Jassy anticipe des milliards d'agents dans les entreprises et les métiers.

[18] NVIDIA — Q1 FY2027 results • 20 mai 2026 (nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2027)

Jensen Huang affirme que l'IA agentique produit déjà du travail et de la valeur à l'échelle des entreprises et des industries.

[19] Prometheus / Jeff Bezos — TechCrunch • 11 juin 2026 (techcrunch.com/2026/06/11/jeff-bezos-prometheus-raises-12b-to-build-an-artificial-general-engineer-for-the-physical-world/)

Prometheus, cofondée par Jeff Bezos et Vik Bajaj, annonce 12 Md\$ de financement à environ 41 Md\$ de valorisation pour une IA orientée ingénierie et monde physique.

[20] Prometheus / Jeff Bezos — Financial Times • 11 juin 2026

Bezos décrit Prometheus comme un projet destiné à transformer ingénierie et fabrication ; source de contre-vérification du financement et du positionnement.

[21] Otto — Hardware • 2026 (myotto.ai/hardware)

Otto One est présenté comme un ordinateur dédié aux agents : identifiants, fichiers et tâches en local, inférence encore dans le cloud.

[22] Otto — Presale • 2026 (myotto.ai/presale)

Otto One est proposé à partir de 599 \$, préchargé avec OpenClaw ; expédition annoncée pour janvier 2027.

[23] Otto — Contact • 2026 (myotto.ai/contact)

The Tokenry, Inc., l'éditeur d'Otto, indique être basé à New York.

[24] Reid Hoffman & Greg Beato — Superagency • 2025 (simonandschuster.com/books/Superagency/Reid-Hoffman/9798893310108)

Hoffman développe l'idée d'une augmentation de la capacité d'action humaine par l'IA.

[25] Mustafa Suleyman & Michael Bhaskar — The Coming Wave • 2023

(penguinrandomhouse.com/books/722674/the-coming-wave-by-mustafa-suleyman-with-michael-bhaskar/)

L'ouvrage analyse la diffusion de systèmes puissants et le problème du contrôle.

[26] Kissinger, Schmidt & Mundie — Genesis • 2024

(hachettebookgroup.com/titles/henry-a-kissinger/genesis/9780316581295)

L'ouvrage interroge les conséquences d'une IA gagnant en autonomie et en capacité d'action.

[27] Légifrance — Article L223-1 du Code de la consommation • 11 août 2026
(legifrance.gouv.fr/codes/article_lc/LEGIARTI000051830285/2026-08-11)

En France, le démarchage téléphonique B2C bascule vers le consentement préalable, que le professionnel doit pouvoir prouver.

[28] Légifrance — Décret n° 2026-662 • 23 juil. 2026 (legifrance.gouv.fr/jorf/id/JORFTEXT000054483939)

Le décret précise notamment, pour certains secteurs, les conditions de rappel d'une demande d'information explicite et le délai de cinq jours ouvrables.

[29] CNIL — Prospection électronique • 2026 (cnil.fr/fr/la-prospection-commerciale-par-courrier-electronique-sms-mms-et-automate-dappel)

La prospection électronique B2C repose en principe sur le consentement préalable, avec exceptions encadrées.

[30] FCC — AI-generated voices in robocalls • 8 févr. 2024 (fcc.gov/document/fcc-confirms-tcpa-applies-ai-technologies-generate-human-voices)

La FCC confirme l'application du TCPA aux voix générées par IA dans les robocalls.

Annexe F — Repères 2017-2026 : comment l’IA moderne s’est construite

Pourquoi cette chronologie : trois trajectoires qui convergent

Les modèles, le RAG et l’agentique ont évolué en moins de dix ans à une vitesse rarement observée dans l’histoire des technologies d’entreprise. Reconstituer cette progression permet de comprendre comment les briques se sont assemblées et pourquoi une architecture durable doit rester flexible, multi-modèles et capable d’intégrer rapidement de nouvelles avancées. Cette chronologie retrace également les briques qui ont progressivement rendu possible la conception de SACHA OMON AI ; elle synthétise les apprentissages et les choix de Salomon Bellaïche au fil de cette évolution technologique.

01 · MODÈLES ET GRAND PUBLIC	02 · RAG ET CONNAISSANCE	03 · AGENTIQUE ET ENTREPRISE
Ce que les modèles savent comprendre, produire et rendre accessible.	Ce que le système peut retrouver, vérifier, citer et actualiser.	Ce que le système peut planifier, transmettre, exécuter et gouverner.

Lecture : chaque intitulé de jalon renvoie à la publication scientifique, à l’annonce officielle ou à la documentation primaire correspondante. La sélection retient les versions et les lancements qui ont modifié l’usage public, le RAG ou le passage en production dans l’entreprise ; elle ne prétend pas lister chaque mise à jour mineure de chaque modèle.

Pourquoi les trois histoires n’avancent pas au même rythme

1. Les modèles occupent l’espace public. Leurs noms et leurs versions sont visibles : GPT, Claude, Gemini, Grok, Llama, Mistral ou DeepSeek. Chaque lancement devient un événement, alors que l’entreprise n’achète jamais seulement un nom de modèle.
2. Le RAG progresse par qualité de connaissance. Ses avancées portent sur la recherche dense ou hybride, le classement, la contextualisation, les graphes, l’évaluation et la multimodalité. Elles sont moins médiatiques, mais déterminent la fiabilité de la réponse.
3. L’agentique arrive plus tard dans l’entreprise. Agir exige davantage que générer du texte : identité, droits d’accès, outils, mémoire, reprise par l’humain, observabilité, évaluations et sécurité. Ces couches deviennent réellement industrielles en 2025 et 2026.

La bonne unité de lecture : ne pas compter les annonces, mais observer quand le modèle, la connaissance et la capacité d’action deviennent suffisamment mûrs pour fonctionner ensemble dans une organisation.

PHASE I · LES FONDATIONS

2017-2020 : Transformer, API et mémoire externe

Cette chronologie commence le 12 juin 2017 avec la publication d'« Attention Is All You Need » par une équipe de Google Brain et Google Research. L'intelligence artificielle existait depuis des décennies ; ce travail marque toutefois un point de bascule de l'IA générative moderne, car il introduit l'architecture Transformer, devenue ensuite une fondation de BERT, de la famille GPT et de nombreux grands modèles de langage. Entre 2017 et 2020, trois conditions décisives se rapprochent : une architecture capable de traiter le langage à grande échelle, un accès programmatique par API et une mémoire documentaire externe consultée au moment de répondre.

EN CLAIR · QU'EST-CE QU'UN TRANSFORMER ?

Un Transformer est une architecture de réseau neuronal qui utilise un mécanisme d'attention pour déterminer quelles parties d'un texte ou d'une séquence comptent le plus les unes par rapport aux autres. Par rapport aux architectures récurrentes dominantes à l'époque, il permet de traiter beaucoup plus de relations en parallèle pendant l'entraînement. Cette capacité facilite le passage à l'échelle et la prise en compte du contexte : elle est devenue l'une des bases techniques des LLM modernes.

Document original : « Attention Is All You Need », arXiv, 12 juin 2017, Google Research · ChatGPT, OpenAI, 30 novembre 2022.

01 · MODÈLES ET GRAND PUBLIC

Juin 2017 Google Brain et Google Research : Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser et Illia Polosukhin présentent le Transformer, une architecture fondée sur l'attention. La diffusion ouverte de ce travail permet au Transformer de devenir le socle commun des grandes familles de modèles qui suivront.

Juin 2018 OpenAI : premier GPT

Le préentraînement génératif suivi d'une adaptation à plusieurs tâches ouvre la lignée GPT.

Oct. 2018 Google Research : BERT

La compréhension bidirectionnelle issue des Transformers change la recherche d'information et la représentation du langage.

Fév. 2019 OpenAI : GPT-2

La génération cohérente sans entraînement propre à chaque tâche rend visible le potentiel des modèles généralistes.

Mai 2020 OpenAI : GPT-3

Le changement d'échelle montre qu'un modèle peut accomplir de nombreuses tâches à partir d'instructions et de quelques exemples.

02 · RAG ET CONNAISSANCE

Avr. 2020 Meta AI : Dense Passage Retrieval

La recherche dense apprend à rapprocher une question et des passages pertinents, brique déterminante des futurs systèmes RAG.

Mai 2020 Lewis et al. : Retrieval-Augmented Generation

Le RAG formalise l'association entre la mémoire paramétrique du modèle et une mémoire documentaire externe recherchée au moment de la réponse.

03 • AGENTIQUE ET ENTREPRISE**Juin 2020 OpenAI : ouverture de l'API**

L'accès programmatique fait entrer les modèles dans les produits et les systèmes d'entreprise : le modèle devient une couche callable, et non plus seulement un objet de recherche.

CE QUE L'ENTREPRISE DOIT RETENIR

Entre 2017 et 2020, les briques essentielles se mettent en place : un modèle qui comprend le contexte, une connaissance externe qu'il peut consulter et un accès par API qui permet de l'intégrer aux applications. L'agent complet n'existe pas encore, mais ses dépendances fondamentales sont déjà réunies.

PHASE II • L'INSTRUCTION, LA RECHERCHE ET L'ACTION**2021-2022 : le modèle apprend à chercher, suivre une intention et agir**

Les trois fils commencent à se rapprocher. Le modèle suit mieux les intentions humaines, la recherche est intégrée à la réponse et les premiers cadres de raisonnement-action décrivent explicitement une boucle agentique.

01 • MODÈLES ET GRAND PUBLIC**Janv. 2022 OpenAI : InstructGPT**

L'entraînement avec retour humain améliore le suivi des instructions et prépare la forme conversationnelle qui sera adoptée à grande échelle.

30 nov. 2022 OpenAI : lancement public de ChatGPT

L'interface conversationnelle rend l'IA générative compréhensible par le grand public et accélère brutalement la demande des entreprises.

02 • RAG ET CONNAISSANCE**Déc. 2021 OpenAI : WebGPT**

Le modèle apprend à utiliser un navigateur, consulter des sources et produire des réponses avec références : un pont entre recherche, action et génération documentée.

03 • AGENTIQUE ET ENTREPRISE**Mai 2022 MRKL Systems : routage entre modèle et modules externes**

Le modèle est présenté comme un routeur capable de solliciter calculatrices, bases et services spécialisés plutôt que de tout produire seul.

Oct. 2022 Yao et al. : ReAct

Le modèle alterne raisonnement, action et observation dans une boucle ; cette structure devient l'une des fondations les plus claires de l'agentique moderne.

CE QUE L'ENTREPRISE DOIT RETENIR

ChatGPT révèle la puissance de l'interface conversationnelle ; ReAct et MRKL montrent ce qu'il lui manque pour l'entreprise : choisir une action, utiliser un outil, observer le résultat et recommencer sous contrôle.

PHASE III • L'ÉCOSYSTÈME S'OUVRE**2023 : plusieurs familles de LLM, des outils et les premiers services d'entreprise**

L'année 2023 ne se résume pas à GPT-4. Claude, Llama, Mistral et Gemini installent un marché multi-modèles, pendant que le RAG et l'appel d'outils deviennent des fonctions accessibles aux développeurs et aux entreprises.

01 • MODÈLES ET GRAND PUBLIC**24 fév. 2023 Meta : LLaMA**

Une famille de modèles ouverts accélère la recherche et la possibilité de déployer des modèles dans des environnements maîtrisés.

14 mars 2023 OpenAI : GPT-4

GPT-4 apporte des capacités multimodales et une meilleure exécution d'instructions complexes dans ChatGPT et par API.

14 mars 2023 Anthropic : Claude

Claude devient accessible par interface de conversation et par API, créant une deuxième grande famille propriétaire pour les organisations.

18 juil. 2023 Meta et Microsoft : Llama 2

La licence autorise la recherche et l'usage commercial, élargissant fortement le choix des entreprises.

27 sept. 2023 Mistral AI : Mistral 7B

Un acteur européen propose un modèle compact sous licence Apache 2.0, important pour le coût, la souveraineté et le déploiement local.

6 déc. 2023 Google DeepMind : Gemini 1.0

Gemini est conçu nativement comme une famille multimodale, du terminal mobile aux plus grandes charges de travail.

02 • RAG ET CONNAISSANCE**13 sept. 2023 AWS : aperçu des Agents et Knowledge Bases for Amazon Bedrock**

AWS assemble modèles, sources d'entreprise, embeddings, base vectorielle et RAG dans un service cloud intégré.

17 oct. 2023 Self-RAG

La récupération devient adaptative : le modèle peut décider quand rechercher, évaluer les passages et critiquer sa propre production.

28 nov. 2023 AWS : Knowledge Bases for Amazon Bedrock disponible

Le RAG géré devient un service d'entreprise, relié à des modèles et à des mécanismes d'orchestration.

03 • AGENTIQUE ET ENTREPRISE

9 fév. 2023 Toolformer

Le modèle apprend à choisir et appeler des outils externes par API, capacité centrale d'un agent capable d'agir.

13 juin 2023 OpenAI : function calling

Les modèles peuvent produire des appels structurés vers les fonctions d'une application, ce qui relie conversation et systèmes métiers.

25 sept. 2023 Microsoft Research : AutoGen

Un framework ouvert permet de composer plusieurs agents, humains et outils dans des conversations coordonnées.

28 nov. 2023 AWS : Agents for Amazon Bedrock

Le service décompose une demande en étapes, appelle des API d'entreprise et consulte les Knowledge Bases par RAG.

CE QUE L'ENTREPRISE DOIT RETENIR

Dès 2023, l'entreprise peut choisir plusieurs familles de modèles, les relier à ses documents et appeler ses propres fonctions. C'est le moment où le prototype conversationnel commence à devenir une architecture métier.

PHASE IV • LA CONVERGENCE

2024 : multimodalité, RAG correctif, graphes et premiers standards

En 2024, les modèles voient, entendent et raisonnent mieux. En parallèle, le RAG se corrige, se contextualise et s'étend aux graphes. L'entreprise découvre les agents reliés au CRM, aux données et aux outils réels.

01 • MODÈLES ET GRAND PUBLIC

4 mars 2024 Anthropic : Claude 3

Les modèles Haiku, Sonnet et Opus apportent la vision et différents compromis de vitesse, coût et capacité.

18 avr. 2024 Meta : Llama 3

L'écosystème ouvert progresse en raisonnement, code et compréhension multilingue.

13 mai 2024 OpenAI : GPT-4o

Le même modèle traite texte, vision et audio en temps réel, ce qui rapproche l'IA du web, du téléphone et des interfaces vocales.

21 juin 2024 Anthropic : Claude 3.5 Sonnet

Le modèle améliore le raisonnement, la vision et l'exécution de tâches de code à un coût intermédiaire.

23 juil. 2024 Meta : Llama 3.1 405B

Un modèle ouvert de très grande taille renforce la possibilité de choisir entre services propriétaires et déploiements contrôlés.

13 août 2024 xAI : Grok-2

xAI installe une famille concurrente orientée conversation, code et raisonnement, accessible ensuite par API.

12 sept. 2024 OpenAI : o1-preview

Une famille dédiée au raisonnement approfondi montre l'intérêt du temps de calcul au moment de répondre.

11 déc. 2024 Google : Gemini 2.0

Google présente un modèle conçu pour l'ère agentique, avec usage natif d'outils et sorties multimodales.

02 • RAG ET CONNAISSANCE**29 janv. 2024 CRAG : Corrective Retrieval-Augmented Generation**

Un évaluateur juge la qualité des documents retrouvés et déclenche des stratégies correctives lorsque la récupération est insuffisante.

Avr. 2024 Microsoft Research : GraphRAG

Les graphes de connaissances dérivés des documents permettent de répondre à des questions globales qui échappent souvent au découpage classique en fragments.

19 sept. 2024 Anthropic : Contextual Retrieval

Chaque fragment reçoit un contexte propre au document avant indexation, afin d'améliorer la pertinence de la recherche hybride.

16 déc. 2024 Microsoft : GraphRAG 1.0

GraphRAG franchit le seuil d'une version structurée pour les développeurs et les cas complexes de synthèse documentaire.

03 • AGENTIQUE ET ENTREPRISE**17 sept. 2024 Salesforce : lancement d'Agentforce**

Les agents entrent dans les ventes, le service, le marketing et le commerce, reliés au CRM et à Data Cloud.

25 nov. 2024 Anthropic : Model Context Protocol (MCP)

Un protocole ouvert standardise la connexion entre assistants, données et outils externes.

19 déc. 2024 Anthropic : Building Effective Agents

Le modèle augmenté par la récupération, les outils et la mémoire est posé comme le bloc élémentaire d'un système agentique.

CE QUE L'ENTREPRISE DOIT RETENIR

L'année 2024 transforme la qualité du contexte : recherche corrective, graphes, contextualisation des fragments et premiers standards de connexion. L'agent devient crédible parce qu'il est mieux documenté et mieux relié aux systèmes.

PHASE V • PASSAGE À LA PRODUCTION**2025, janvier-juin : raisonner, rechercher et agir dans les systèmes réels**

Au premier semestre 2025, les modèles de raisonnement se multiplient. Les grands fournisseurs livrent des SDK, des protocoles et des moteurs RAG gérés : l'enjeu devient moins le chatbot que l'architecture complète.

01 • MODÈLES ET GRAND PUBLIC**20 janv. 2025 DeepSeek : R1**

Un modèle de raisonnement ouvert, accompagné de modèles distillés, élargit fortement la compétition sur le prix, le déploiement et la recherche.

19 fév. 2025 xAI : Grok 3

xAI fait évoluer Grok vers le raisonnement avancé et prépare des usages d'agents dotés d'outils.

24 fév. 2025 Anthropic : Claude 3.7 Sonnet

Le raisonnement hybride permet de choisir entre réponse rapide et réflexion prolongée dans un même modèle.

25 mars 2025 Google : Gemini 2.5

Google introduit une famille de modèles de raisonnement conçue pour les tâches complexes et multimodales.

5 avr. 2025 Meta : Llama 4

La nouvelle génération ouverte devient nativement multimodale et adopte une architecture mixture-of-experts.

16 avr. 2025 OpenAI : o3 et o4-mini

Les modèles de raisonnement utilisent directement des outils dans ChatGPT, dont recherche web, fichiers, Python et vision.

22 mai 2025 Anthropic : Claude 4

Opus 4 et Sonnet 4 visent explicitement le code, le raisonnement avancé et les agents.

02 · RAG ET CONNAISSANCE**15 janv. 2025 Google Cloud : Vertex AI RAG Engine**

Le cycle ingestion, découpage, indexation, récupération et génération devient un service managé pour les développeurs d'entreprise.

13 juin 2025 Anthropic : système de recherche multi-agent

La recherche dynamique multi-étapes est distinguée du RAG statique : l'agent adapte ses recherches aux éléments trouvés et synthétise les résultats.

03 · AGENTIQUE ET ENTREPRISE**11 mars 2025 OpenAI : Responses API, outils et Agents SDK**

Recherche web, recherche dans les fichiers, usage d'ordinateur et traçage deviennent des composants unifiés pour construire des agents.

9 avr. 2025 Google : Agent2Agent Protocol (A2A)

A2A propose un langage ouvert pour permettre à des agents de fournisseurs différents de se découvrir et de collaborer.

27 mai 2025 Mistral : Agents API

Mistral fournit une API d'agents reliés à des connecteurs, des outils, du code et des bibliothèques documentaires.

CE QUE L'ENTREPRISE DOIT RETENIR

Au premier semestre 2025, la frontière entre modèle, RAG et agent se déplace. Les modèles raisonnent et utilisent des outils ; les plateformes exposent recherche et traçage ; les protocoles commencent à rendre les agents interopérables.

PHASE VI • LES PLATEFORMES AGENTIQUES

2025, juillet-décembre : modèles, connaissance et gouvernance se rejoignent

Le second semestre 2025 voit apparaître les plateformes complètes : environnements d'exécution, mémoire, identité, outils, évaluations, observabilité et politiques. Le modèle devient une brique remplaçable à l'intérieur d'un système plus vaste.

01 • MODÈLES ET GRAND PUBLIC

9 juil. 2025 xAI : Grok 4

Grok 4 intègre l'usage natif d'outils et la recherche en temps réel, caractéristiques directement utiles aux agents.

5 août 2025 OpenAI : gpt-oss

OpenAI publie deux modèles à poids ouverts orientés raisonnement et usage d'outils, adaptés à des déploiements contrôlés.

7 août 2025 OpenAI : GPT-5

GPT-5 unifie davantage réponse rapide, raisonnement et choix d'outils dans ChatGPT et l'API.

18 nov. 2025 Google : Gemini 3

La nouvelle génération met l'accent sur le raisonnement, la multimodalité et l'exécution de tâches complexes.

1^{er} déc. 2025 DeepSeek : V3.2

DeepSeek poursuit la convergence entre raisonnement, génération générale et capacités utiles aux agents.

2 déc. 2025 Mistral : Mistral 3

Mistral Large 3 et plusieurs modèles compacts sous licence Apache 2.0 renforcent le choix entre cloud, privé et local.

02 • RAG ET CONNAISSANCE

30 juil. 2025 Google : Gemini Embedding

Les embeddings sont présentés comme une fondation du RAG et de l'ingénierie du contexte pour les futurs agents.

14 août 2025 AWS : agentic RAG dans Amazon Q Business

La récupération devient un parcours planifié : l'agent choisit plusieurs stratégies de navigation au lieu d'une seule recherche statique.

29 sept. 2025 Anthropic : context engineering pour les agents

La qualité du système dépend de la sélection et du maintien du contexte, pas seulement du texte d'un prompt.

03 • AGENTIQUE ET ENTREPRISE

16 juil. 2025 AWS : aperçu d'Amazon Bedrock AgentCore

AWS fournit une infrastructure modulaire pour exécuter des agents avec mémoire, identité, passerelle, outils et observabilité.

17 sept. 2025 Salesforce : premier Agentic Enterprise Index

Les données d'usage publiées par Salesforce montrent que les agents passent des prototypes aux conversations de service et aux opérations réelles.

6 oct. 2025 OpenAI : AgentKit

OpenAI regroupe conception visuelle, connecteurs, évaluations et déploiement des interfaces d'agents.

13 oct. 2025 AWS : AgentCore disponible

Le runtime, la mémoire, l'identité et l'observabilité agentiques deviennent des services de production avec réseau privé et infrastructure déclarative.

9 déc. 2025 Anthropic : MCP rejoint l'Agentic AI Foundation

MCP devient un projet de fondation ouverte ; Anthropic indique alors plus de 10 000 serveurs publics actifs et une adoption multi-fournisseurs.

CE QUE L'ENTREPRISE DOIT RETENIR

Le second semestre 2025 marque le véritable passage à la production : exécution isolée, mémoire, identité, observabilité, évaluation et politiques rejoignent les modèles et la connaissance. L'architecture compte désormais davantage que le classement d'un LLM.

PHASE VII • INDUSTRIALISER SOUS CONTRÔLE**2026, janvier-avril : identité, politiques, évaluations et RAG multimodal**

Au début de 2026, le sujet n'est plus seulement de rendre un agent capable. Il faut limiter ses droits, observer ses parcours, évaluer la qualité et lui donner accès à une connaissance multimodale : texte, image, audio, vidéo et documents.

01 • MODÈLES ET GRAND PUBLIC**5 fév. 2026 Anthropic : Claude Opus 4.6**

Le modèle soutient des tâches agentiques plus longues et apporte une fenêtre de contexte d'un million de tokens en bêta.

17 fév. 2026 Anthropic : Claude Sonnet 4.6

Le raisonnement long, l'usage d'ordinateur et la planification d'agents deviennent accessibles dans une gamme plus économique.

19 fév. 2026 Google : Gemini 3.1 Pro

Google améliore le raisonnement complexe et l'usage d'outils dans sa famille Pro.

5 mars 2026 OpenAI : GPT-5.4

Le modèle vise les travaux professionnels longs et structurés, avec usage d'outils et analyse de documents.

23 avr. 2026 OpenAI : GPT-5.5

La nouvelle génération améliore les tâches de connaissance, la programmation et les parcours agentiques prolongés.

24 avr. 2026 DeepSeek : V4 Preview

DeepSeek expérimente une nouvelle génération de modèles à très long contexte destinée au raisonnement et aux usages d'agents.

02 • RAG ET CONNAISSANCE**10 mars 2026 Google : Gemini Embedding 2 en aperçu**

Un même espace vectoriel relie texte, images, audio, vidéo et documents, préparant un RAG réellement multimodal.

6 avr. 2026 AWS : recherche hybride pour un assistant agentique

La recherche sémantique et la recherche lexicale sont combinées, puis exploitées dans une boucle agentique avec contrôle de qualité.

30 avr. 2026 Google : Gemini Embedding 2 disponible

Le RAG agentique peut retrouver et rapprocher des contenus dans plus de 100 langues à travers plusieurs modalités.

03 • AGENTIQUE ET ENTREPRISE**13 janv. 2026 Anthropic Labs : Cowork, Skills et MCP**

Anthropic regroupe des prototypes destinés à faire de Claude un collaborateur relié aux outils et aux connaissances du travail.

3 mars 2026 AWS : AgentCore Policy disponible

Les droits d'un agent sur ses outils sont évalués par une politique externe au code de l'agent, avec contrôles fins et traçables.

11 mars 2026 OpenAI : environnement informatique dans Responses API

Les agents reçoivent un environnement d'exécution isolé pour manipuler fichiers, code et outils tout en conservant une trace des opérations.

31 mars 2026 AWS : AgentCore Evaluations disponible

La qualité peut être mesurée en continu sur les traces de production et lors des changements de version.

15 avr. 2026 OpenAI : nouvelle évolution de l'Agents SDK

Le SDK consolide exécution, outils, garde-fous, mémoire de contexte et orchestration de plusieurs agents.

22 avr. 2026 Google Cloud : Gemini Enterprise Agent Platform

La plateforme rassemble modèles, création, orchestration, identité, sécurité, observabilité et gouvernance des flottes d'agents.

CE QUE L'ENTREPRISE DOIT RETENIR

Début 2026, la sécurité et l'évaluation ne sont plus des annexes. Elles deviennent des services distincts qui encadrent chaque appel d'outil. En parallèle, le RAG sort du texte seul et commence à comprendre l'ensemble des médias de l'entreprise.

PHASE VIII • ÉTAT DE L'ART AU 19 AOÛT 2026**2026, mai-août : modèles rapides, agents managés et standards interopérables**

À la date de clôture de ce livre blanc, les nouveautés les plus importantes ne sont plus de simples records de modèle. Elles réduisent le coût des boucles agentiques, rendent les outils interchangeables, renforcent les contrôles et rapprochent l'agent des canaux réels de l'entreprise.

01 • MODÈLES ET GRAND PUBLIC**19 mai 2026 Google : Gemini 3.5 Flash**

Un modèle rapide conçu pour le code et les tâches agentiques devient le moteur par défaut d'expériences grand public comme AI Mode.

22 mai 2026 Mistral : Medium 3.5 et Work mode

Mistral associe un modèle intermédiaire à des agents distants, des connecteurs MCP et des contrôles humains pour le travail en entreprise.

9 juin 2026 Anthropic : Claude Fable 5 et Mythos 5

La famille Claude franchit un nouveau seuil de raisonnement, d'endurance agentique et d'usage spécialisé.

26 juin 2026 OpenAI : aperçu de GPT-5.6 Sol

OpenAI présente une nouvelle génération orientée recherche, code et résolution de problèmes complexes.

30 juin 2026 Anthropic : Claude Sonnet 5

Sonnet 5 est présenté comme le Sonnet le plus agentique, capable de planifier et d'utiliser navigateur et terminal.

9 juil. 2026 OpenAI : GPT-5.6

Le modèle généralise les progrès de Sol et vise des tâches de recherche, de programmation et de connaissance de haut niveau.

21 juil. 2026 Google : Gemini 3.6 Flash

Le modèle réduit les tokens et les appels d'outils nécessaires aux boucles agentiques, à un prix inférieur à la génération précédente.

24 juil. 2026 Anthropic : Claude Opus 5

Opus 5 améliore les tâches longues de connaissance et d'ingénierie et vise une exécution agentique plus stable.

6 août 2026 OpenAI : GPT-5.6 Sol et Luna dans ChatGPT

Sol progresse pour les offres Plus et Pro ; Luna devient le modèle par défaut des offres Free et Go, montrant la diffusion rapide des nouvelles générations vers le grand public.

02 • RAG ET CONNAISSANCE**16 juil. 2026 AWS : Managed Knowledge Base pour la recherche d'entreprise**

AWS clarifie l'éventail entre base de connaissances entièrement gérée et pipeline RAG construit sur mesure pour les agents.

28 juil. 2026 MCP, spécification 2026-07-28

La spécification adopte un cœur sans session, des résultats cachables et des extensions pour les interfaces et les tâches longues, afin de passer à l'échelle cloud.

03 • AGENTIQUE ET ENTREPRISE**19 mai 2026 Google : Managed Agents dans Gemini API**

Un appel d'API peut lancer un agent qui raisonne, utilise des outils et exécute du code dans un environnement Linux isolé.

17 juin 2026 Google : Agentic Resource Discovery

Une spécification ouverte vise la découverte et la vérification d'outils, de compétences et d'agents à travers le web.

17 juin 2026 AWS : Guardrails dans AgentCore Policy

Les politiques d'accès aux outils sont complétées par des protections contre l'injection de prompt et l'exposition de données sensibles.

22 juil. 2026 OpenAI : Presence pour les agents d'entreprise

OpenAI présente une couche destinée à déployer et superviser des agents reliés aux connaissances et aux outils internes.

28 juil. 2026 Google : Managed Agents avec contrôles de budget et audits d'outils

Les environnements managés ajoutent des hooks de contrôle, des plafonds de tokens et des déclencheurs planifiés.

30 juil. 2026 AWS : Bedrock Agents Classic passe en maintenance

AWS oriente les nouveaux projets vers AgentCore, signe du remplacement des premiers agents gérés par une plateforme de production plus complète.

6 août 2026 Google et partenaires : Agent Plugins 1.0

Un format indépendant des fournisseurs regroupe compétences, instructions et serveurs MCP dans des modules portables.

7 août 2026 Salesforce : Agentic Enterprise Index 2025-2026

Salesforce publie une hausse proche de trois fois du nombre d'agents activés parmi les organisations étudiées et documente leur passage à l'échelle.

CE QUE L'ENTREPRISE DOIT RETENIR

Au 19 août 2026, le marché converge vers des agents managés, multi-modèles et interopérables. La valeur durable ne vient plus d'un fournisseur unique : elle vient de la connaissance propriétaire, des règles, des canaux et de la qualité du relais entre l'IA et les équipes humaines.

Mentions légales

Éditeur

SACHA OMON AI, société de droit israélien. Directeur de la publication et rédacteur en chef : Salomon Bellaïche. Contact : sacha@sacha-omon.com · www.sacha-omon.ai.

Droits d’auteur

Ce livre blanc est protégé par le droit d’auteur israélien (Copyright Act, 5768-2007) et, par la Convention de Berne, dans tous les pays où il sera lu. La protection existe du seul fait de sa création : aucun dépôt n’est nécessaire. Elle couvre le texte, les titres, les tableaux, les schémas, la structure et la mise en page.

Ce que vous pouvez faire, et ce qui est interdit

AUTORISÉ	INTERDIT SANS ACCORD ÉCRIT
Lire, imprimer et partager le fichier tel quel.	Recopier le document, en totalité ou en grande partie.
Citer des extraits courts en indiquant l’auteur et www.sacha-omon.ai .	Le republier sur un autre site ou un autre support.
Faire un lien vers le document.	Le modifier, le traduire ou le vendre.
Le citer comme référence dans un travail de recherche ou de presse.	En retirer le nom de l’auteur ou la source.

Tableau 58 : Conditions d’utilisation. Toute demande d’autorisation : sacha@sacha-omon.com.

Les moteurs de recherche et les systèmes d’intelligence artificielle peuvent indexer ce document et le citer, à condition d’en indiquer la source.

Sources, marques et responsabilité

Les chiffres et études cités proviennent de leurs éditeurs d’origine, identifiés en annexe, et n’engagent qu’eux. Les informations sont arrêtées au 19 août 2026. Les marques citées appartiennent à leurs détenteurs et leur mention n’implique aucun partenariat.

Ce document est informatif. Il ne constitue ni un conseil juridique, financier ou d’investissement, ni un engagement contractuel, ni une garantie de résultat. Les décisions prises à sa lecture relèvent du lecteur.

© 2026 SACHA OMON AI. Tous droits réservés. Édition internationale 2026, édition n° 1 · version française, relecture n° 14 du 21 août 2026.

Après le Livre blanc, passez à l'action

Vous savez désormais pourquoi l'agentique, le RAG et le multicanal changent la relation commerciale. La question est maintenant simple : où commencer dans votre entreprise ?

DEMANDEZ VOTRE ÉTUDE STRATÉGIQUE

Nous analysons vos canaux, votre parcours commercial, vos volumes de prospects, votre délai de première réponse et votre potentiel de ROI, puis nous formulons une recommandation concrète sur votre premier cas d'usage.

FAITES VOTRE CAHIER DES CHARGES AVEC VOTRE PROPRE IA

Le script autonome des pages 88 à 92 pose huit questions, une par une, puis prépare une première version de votre cahier des charges : les 14 décisions à documenter, les 8 marches appliquées à votre cas, les hypothèses économiques et les questions à poser à un fournisseur.

Copiez le script dans ChatGPT, Claude, Gemini ou Grok. Dix à quinze minutes suffisent pour obtenir une première base de travail à corriger avec vos équipes.

Agent SDR IA agentique, RAG dédié, environnement mono-locataire, plus de 40 langues, 24 heures sur 24 et 7 jours sur 7, cinq canaux : web, WhatsApp, téléphone, e-mail et SMS. L'IA prépare, qualifie et transmet. L'humain garde le dernier mot.

www.sacha-omon.ai

Engagez, convertissez et transformez vos prospects entrants en clients avec Sacha.

Sacha se connecte à votre entreprise, apprend vos offres, vos règles et vos connaissances, et échange instantanément avec vos prospects sur [Web](#), [WhatsApp](#) et [Téléphone](#) — 24/7, en 40+ langues.



Réponse instantanée
en moins de 2 secondes



Qualification des
prospects automatique



Répond sur Web,
WhatsApp et Téléphone



Réserve des rendez-vous
qualifiés pour votre équipe



Disponible 24/7
en 40+ langues

Sacha vous aide à capter, qualifier et convertir plus rapidement la demande entrante. Elle répond aux questions, identifie les véritables intentions d'achat, traite les objections et guide vos prospects vers l'étape suivante — pour que votre équipe puisse se concentrer sur la conclusion des ventes.



WEB



WHATSAPP



TÉLÉPHONE

SE CONNECTE LÀ OÙ SONT VOS PROSPECTS



Facebook



Instagram



Site Web



WhatsApp



Téléphone



**MEILLEURE
CONVERSION**

Convertissez plus de
leads entrants



**QUALIFICATION
PLUS INTELLIGENTE**

Identifiez la véritable
intention d'achat



**PRODUCTIVITÉ
ÉQUIPE**

Concentrez-vous sur
la conclusion des ventes



**SÉCURITÉ
ENTREPRISE**

Vos données,
toujours protégées

Réservez votre démo personnalisée

Découvrez comment Sacha peut qualifier
vos prospects et planifier plus de rendez-vous.

Prendre rendez-vous →



www.sacha-omon.ai



Web



WhatsApp



Téléphone