LES MULTIAGENTS D'IA SOUVERAINS ET ÉVOLUTIFS AU SERVICE DE LA RELATION CLIENT

Chapitre réalisé par

CGI BUSINESS CONSULTING

INTRODUCTION : VERS UNE NOUVELLE ÈRE DES SYSTÈMES D'IA POUR LA RELATION CLIENT

Avec l'essor de l'intelligence artificielle générative, les *chatbots* connaissent un regain d'intérêt dans les stratégies de relation client, alors que les technologies de la génération précédente étaient décevantes. Alimentés par des grands modèles de langage (LLM pour *Large Language Model*) reliés à des bases de connaissances internes via le mécanisme de génération augmentée de récupération (RAG pour *Retrieval-Augmented Generation*), ils ont suscité de grands espoirs en promettant une compréhension plus fine des demandes et des réponses mieux adaptées aux attentes des utilisateurs.

Au-delà du *chatbot* : les limites de l'IA de première génération

Pourtant, dans la pratique, l'enthousiasme initial s'est vite heurté à la réalité du terrain. Malgré des avancées notables par rapport aux anciens systèmes à scripts, ces premiers *chatbots* révèlent vite leurs limites. Leur fonctionnement reste fondamentalement passif : un *chatbot* RAG, par exemple, effectue généralement une seule recherche d'information pour générer sa réponse. Si celle-ci est incomplète ou erronée, l'échange devient vite insatisfaisant, avec des réponses vagues, hors sujet, voire absurdes.

Plus concrètement, ces agents restent figés dans un rôle purement informatif : ils savent extraire des données, mais sont incapables d'agir. Or, comprendre une demande ne suffit plus : encore faut-il pouvoir y répondre efficacement. C'est précisément là que les agents d'IA changent la donne, en apportant cette capacité d'action qui faisait défaut aux premières générations de *chatbots*.

De la parole à l'acte : qu'est-ce qu'un agent d'IA ?

On peut définir un agent d'IA comme une entité autonome capable de poursuivre un objectif en interagissant avec son environnement.

Bien plus qu'un simple interlocuteur, il agit comme un véritable collaborateur numérique. Son « cerveau » — un modèle de langage comme GPT-4o ou Claude Sonnet — lui confère des capacités avancées de raisonnement, de planification et de compréhension du langage.

Toutefois, ce qui fait la force de l'agent d'IA, c'est sa « boîte à outils », soit un ensemble de fonctionnalités qui lui permettent d'intervenir de manière concrète et ciblée dans les processus métiers.

Dans ce contexte, un outil désigne une fonction spécifique que l'agent peut déclencher, comme par exemple :

- Consulter une base de données pour vérifier le statut d'une commande.
- Se connecter à une API externe pour réserver un créneau de livraison.
- Envoyer un e-mail pour confirmer un rendez-vous.

- Mettre à jour une plateforme CRM avec le compte-rendu d'un appel.

Grâce à ces outils, l'agent ne se contente plus de dire « Votre colis est en cours de livraison ». Il peut, de lui-même, interroger le système du transporteur, identifier un retard, et proposer proactivement une nouvelle date au client par courriel, sans aucune intervention humaine.

Du "couteau suisse" à l'équipe d'experts : agent unique et multiagents

Passer de l'information à l'action implique un changement profond dans la manière dont les systèmes sont conçus. La manière dont nous concevons et organisons ces agents d'IA détermine directement leur efficacité, leur fiabilité et leur capacité à gérer des tâches complexes. Deux philosophies s'opposent : le modèle de l'agent unique et celui de l'équipe d'agents.

L'agent unique (*Single-Agent*) est le "couteau suisse" numérique. Un seul "cerveau" (LLM) gère une tâche de bout en bout, avec un jeu d'instructions unique et une mémoire continue. Il tente de tout faire lui-même.

Le système multiagents (*Multi-Agent*) est une "équipe d'experts". Il coordonne plusieurs agents spécialisés, chacun doté d'instructions et d'un contexte limité à sa mission. Pour accomplir une tâche globale, ils collaborent. Cette collaboration est le plus souvent séquentielle, à la manière d'une chaîne de montage : un agent analyse la demande, la transmet à un autre qui vérifie les données, puis à un troisième qui exécute l'action. Plus rarement, la collaboration peut être parallèle. Plusieurs agents travaillent alors simultanément sur différentes facettes d'un problème.

L'architecture multiagents : la spécialisation au service de la fiabilité

Cette distinction architecturale n'est pas qu'un détail technique ; elle est au cœur de l'efficacité des agents dans des processus métier complexes, comme la gestion de la relation client. Si l'idée d'un agent unique ultra-polyvalent est séduisante, c'est l'approche multiagents séquentielle qui se révèle la plus pertinente et robuste pour la plupart des parcours clients.

En réalité, un parcours de relation client ne se résume presque jamais à une action isolée. Il s'articule le plus souvent autour d'une succession d'étapes logiques : comprendre la demande, vérifier les informations, déclencher l'action appropriée, puis informer le client. Confier l'ensemble de ce processus à un seul agent revient à lui demander de tout faire simultanément. Cela représente une charge irréaliste sur le plan technique et très difficile à maintenir dans la durée.

Dans ce cas, le *prompt system*, qui correspond aux instructions de paramétrage d'un agent, deviendrait un enchevêtrement complexe de règles, d'exceptions et de capacités, censé couvrir chaque scénario possible. Le risque d'erreur, d'oubli ou de confusion entre les étapes serait alors bien trop élevé.

L'approche séquentielle multiagents s'inspire directement de notre mode de fonctionnement : chaque agent, tel un membre d'une équipe, intervient au bon moment avec un rôle bien défini. En confiant chaque étape du processus à un agent spécialisé — que ce soit l'analyse de documents, l'interrogation d'une base de données ou la communication avec le client — on construit une chaîne claire, modulaire et robuste.

Cette spécialisation apporte une meilleure lisibilité, une exécution plus fluide et une supervision facilitée. Chaque maillon est identifiable et maîtrisé. C'est précisément cette organisation rigoureuse et prévisible qui permet de faire le pont entre une IA conversationnelle et une IA opérationnelle, capable de piloter des processus métier de bout en bout.

L'IA opérationnelle : levier de productivité et de transformation des métiers

La promesse d'une IA véritablement opérationnelle ne se limite pas à une avancée technologique : elle incarne un changement de fond dans les métiers de la relation client.

Les études les plus récentes convergent vers un même constat. Notamment, le rapport *Future of Jobs 2025* du World Economic Forum identifie la relation client comme l'un des secteurs les plus concernés par l'automatisation.

Les tendances à venir sont claires : d'ici 2028, près de 70 % des interactions de service client pourraient être gérées par des IA agentiques. Certaines grandes entreprises anticipent déjà ce virage. À titre d'exemple, Andy Jassy, CEO d'Amazon, a récemment confirmé que le groupe entrait dans une phase de rationalisation et de décrue des effectifs, avec pour objectif de tirer pleinement parti des gains de productivité liés à l'IA.

Deux piliers clés pour déployer : souveraineté et passage à l'échelle

Si la fiabilité fonctionnelle, sur le plan métier, constitue un prérequis essentiel, elle ne suffit pas, à elle seule, à garantir le succès. Pour qu'une architecture multiagents passe du statut de solution prometteuse à celui d'atout stratégique réellement déployable, deux autres piliers semblent essentiels dans le contexte actuel : la souveraineté et l'évolutivité.

La souveraineté est aujourd'hui un enjeu central. Elle traduit la volonté des entreprises de garder le contrôle sur leurs données, leurs modèles et leurs infrastructures. Alors que la plupart des services d'IA s'appuient encore sur des solutions externes, un nombre croissant d'organisations cherchent désormais à maîtriser l'exécution des agents.

Ce besoin devient d'autant plus pressant dans un contexte géopolitique instable, où les tensions commerciales ravivent les craintes liées à la dépendance technologique. Face à ces incertitudes, les entreprises sont poussées à privilégier des solutions souveraines — ou, à défaut, à anticiper des scénarios de repli sécurisés et maîtrisés.

L'évolutivité désigne la capacité à opérer à grande échelle sans compromettre les performances ni entraîner une hausse excessive des coûts. Idéalement, un système capable de fonctionner pour 1000 utilisateurs doit pouvoir en gérer 10 000 avec la même fluidité, pour absorber un pic de demandes.

C'est cette capacité d'extension maîtrisée — conçue pour être évolutive — qui permet de faire passer une solution d'un simple prototype à un levier stratégique durable pour la relation client.

DU CONCEPT À LA PRATIQUE

LES CAS D'USAGE DES SYSTÈMES MULTIAGENTS CLIENT

Comme nous l'avons souligné précédemment, les LLM utilisés de manière isolée, comme dans un *chatbot* RAG classique, peuvent informer, mais pas agir. La première évolution naturelle consiste à transformer ce LLM en un agent unique en lui fournissant des outils pour exécuter des tâches. Cependant, ce modèle de « couteau suisse » atteint vite ses limites face à la complexité, puisque la « surcharge cognitive » nuit à sa fiabilité.

C'est ici qu'intervient la véritable rupture : l'architecture multiagents.

En pratique, cette approche résout un problème que rencontrent de nombreuses entreprises : la fragmentation. Aujourd'hui, les briques d'IA — FAQ dynamique, *chatbot*, analyseur d'e-mails — coexistent mais opèrent en silos. Incapables de communiquer, elles renvoient une vision fragmentée du client et aboutissent à des interactions incohérentes.

La promesse des systèmes multiagents (SMA) est de transformer cet empilement d'outils en une « équipe d'experts » virtuelle et coordonnée. Grâce à un orchestrateur central, qui agit comme un véritable coordinateur, l'entreprise peut enfin :

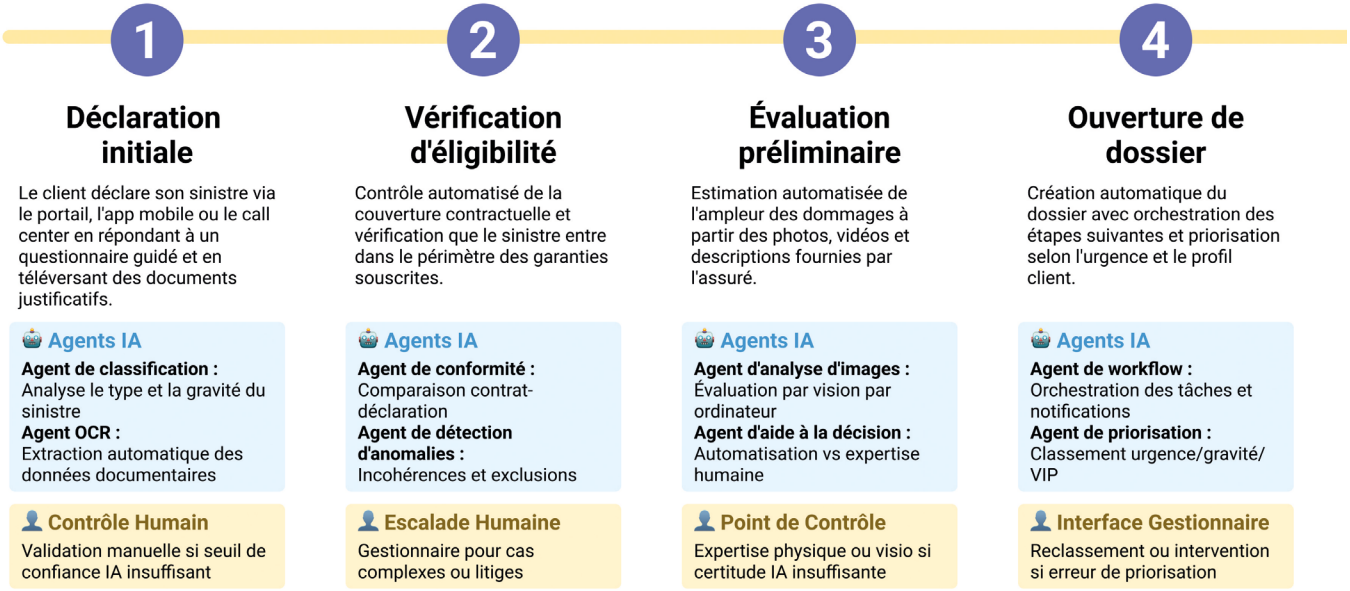
1. Briser les silos en faisant collaborer les agents spécialisés.
2. Garantir sa souveraineté en s'assurant que les processus suivent des règles métier précises et contrôlées.
3. Passer à l'échelle en automatisant les tâches séquentielles et en ne sollicitant l'intervention humaine que lorsque sa valeur ajoutée est maximale.

Pour mieux comprendre la portée concrète de cette approche, explorons à présent un parcours client typique dans le secteur de l'assurance. Le scénario suivant illustre comment cette « chaîne de montage » intelligente, pilotée par un orchestrateur, fluidifie l'expérience client.

Lors de la première étape du processus — la déclaration — le client transmet un constat accompagné de photos. Un premier agent de catégorisation identifie qu'il s'agit d'un sinistre automobile. Un second agent, doté de capacités OCR (*Optical Characters Recognition*), lit les documents, tandis qu'un agent d'extraction isole les données clés : numéro de contrat, date, immatriculation, etc.

Une fois ces données consolidées, un agent de conformité vérifie la couverture du contrat. Simultanément, un agent de détection d'anomalies recherche d'éventuelles fraudes. Si tout est en ordre, un agent d'analyse d'images estime les dommages.

Sur cette base, un agent de décision valide le dossier si le cas est simple, ou le transmet à un expert humain si une incertitude subsiste.



Une fois la décision prise, un agent d'indemnisation calcule le remboursement. Enfin, un agent de communication notifie le client et un agent de satisfaction recueille son *feedback*.

Ce parcours, entièrement orchestré, démontre comment l'IA peut gérer un processus complexe avec rigueur et réactivité, en ne sollicitant l'humain que lorsque son jugement est indispensable.

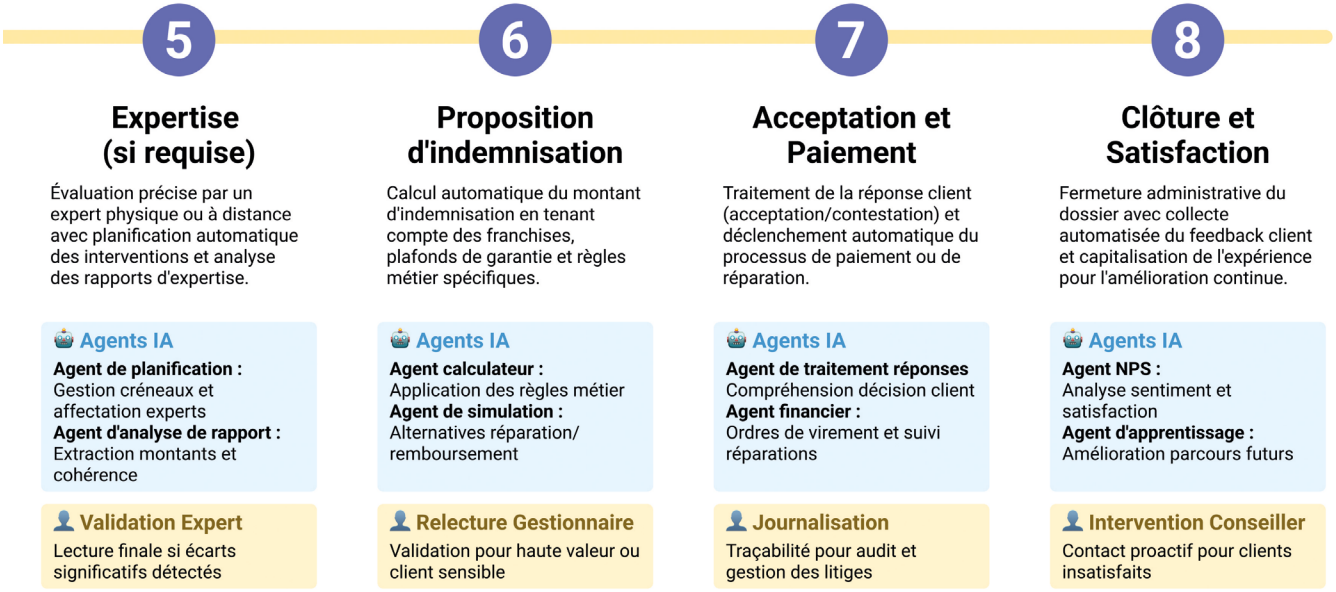
CINQ CAS D'USAGE OÙ LES SYSTÈMES MULTIAGENTS REDÉFINISSENT L'EXPÉRIENCE CLIENT

La grande force des systèmes multiagents réside dans leur capacité à gérer des processus complexes de bout en bout. Chaque agent y remplit une mission spécifique. Leur collaboration est orchestrée pour garantir un fonctionnement harmonieux, que ce soit au sein d'une chaîne séquentielle ou via un coordinateur central. Voici des cas d'usage concrets illustrant cette approche.

1. Vente assistée et recommandations personnalisées

Dans le cadre du e-commerce, un système multiagents peut analyser en continu les comportements de navigation pour générer des recommandations dynamiques. Un agent de profilage étudie le parcours, un agent de recommandation propose des produits, un agent de configuration gère la compatibilité d'éléments complexes (comme un PC sur mesure), et un agent de tarification ajuste les prix.

Le groupe H&M, géant suédois de la mode accessible, a intégré cette approche dans le cadre de



sa stratégie de transformation digitale. Engagé de manière croissante en faveur de la durabilité, il a mis en place un assistant digital reposant sur des agents spécialisés. Cette initiative a permis d'enregistrer une hausse de 25 % des conversions sur les parcours assistés.

Source : https://redresscompliance.com/how-hm-uses-ai-powered-chatbots-to-improve-customer-service/

2. Support en temps réel pour les conseillers « augmentés »

Dans les centres d'appel, plusieurs agents d'IA peuvent aider un conseiller humain en temps réel. Lors d'un échange avec un client, plusieurs agents interviennent simultanément pour enrichir la qualité du service, tout en préservant la fluidité du dialogue.

Un agent de transcription convertit la parole en texte en temps réel, un agent d'assistance propose des suggestions de réponse ou des fiches pratiques, un agent de détection d'émotions identifie les signes de stress ou d'insatisfaction dans la voix, et un agent de rédaction génère automatiquement le résumé post-appel et un courriel de suivi.

Cette approche a été mise en œuvre chez Allianz Direct, la filiale 100 % digitale du groupe Allianz. Pour réduire la charge cognitive liée à la recherche documentaire, un assistant intelligent a été déployé afin de soutenir les conseillers dans leurs tâches quotidiennes. Ce dispositif a permis d'améliorer leur confort de travail tout en augmentant la précision des réponses de 10 à 15 %.

Source : https://www.zenml.io/llmops-database/rag-powered-agent-assist-tool-for-insurance-contact-centers

3. Gestion proactive de la relation

Dans les secteurs où les taux de résiliation sont élevés, leur anticipation est devenue un enjeu stratégique majeur. Les systèmes multiagents permettent de renforcer la qualité des démarches proactives d'analyse et d'activation.

Ils permettent de détecter les signaux faibles de désengagement et de déclencher une réponse ciblée, avant même que le client n'exprime son insatisfaction.

Un agent d'analyse comportementale identifie les baisses d'usage ou les interactions anormales, un agent prédictif attribue un score d'attrition (*churn*) en croisant historique et contexte, et un agent de génération de solution propose automatiquement une offre de rétention personnalisée : geste commercial, avantage tarifaire ou appel prioritaire.

Deutsche Telekom, l'un des principaux opérateurs télécoms mondiaux, a automatisé à grande échelle sa relation client en s'appuyant sur une plateforme multiagents déployée dans 10 pays. Résultat : 60 % des demandes sont désormais traitées sans intervention humaine, entraînant une amélioration des performances de +38 % et un net gain en satisfaction client.

Source : https://www.zenml.io/llmops-database/building-a-multi-agent-llm-platform-for-customer-service-automation

4. Automatisation du parcours contractuel

Dans certains secteurs (banque, assurance, crédit, B2B…), les processus contractuels sont particulièrement longs, fragmentés et soumis à de multiples validations.

Dans ce contexte, les systèmes multiagents offrent une réponse efficace, en permettant d'orchestrer l'ensemble du cycle de vie d'un contrat : depuis la prise de contact initiale, jusqu'à la signature électronique et l'*onboarding*.

Un agent classificateur d'intention détecte la demande de devis, un agent de génération rédige un brouillon, un agent de vérification relit les clauses, et un agent d'*onboarding* déclenche les étapes post-signature.

La plateforme SaaS Vendr, spécialisée dans l'optimisation des achats logiciels, a ainsi industrialisé l'extraction de données sur plus de 100 000 documents contractuels grâce à une chaîne multiagents, garantissant une haute précision.

Source : https://www.zenml.io/llmops-database/scaling-document-processing-with-llms-and-human-review

5. De l'optimisation de parcours à l'intégration totale : le cas Klarna

Ces cas d'usage illustrent la capacité des architectures multiagents à piloter un parcours client spécifique avec une grande efficacité.

La prochaine étape, déjà franchie par certains, consiste à passer de ces parcours « verticaux » à une orchestration plus horizontale et transversale. L'enjeu n'est plus seulement d'optimiser une étape, mais d'intégrer un processus entier de la relation client.

C'est précisément ce qu'a mis en place l'entreprise klarna. Klarna, *leader* mondial des services de paiement et de financement, faisait face à un défi majeur : plus de 2,3 millions de conversations

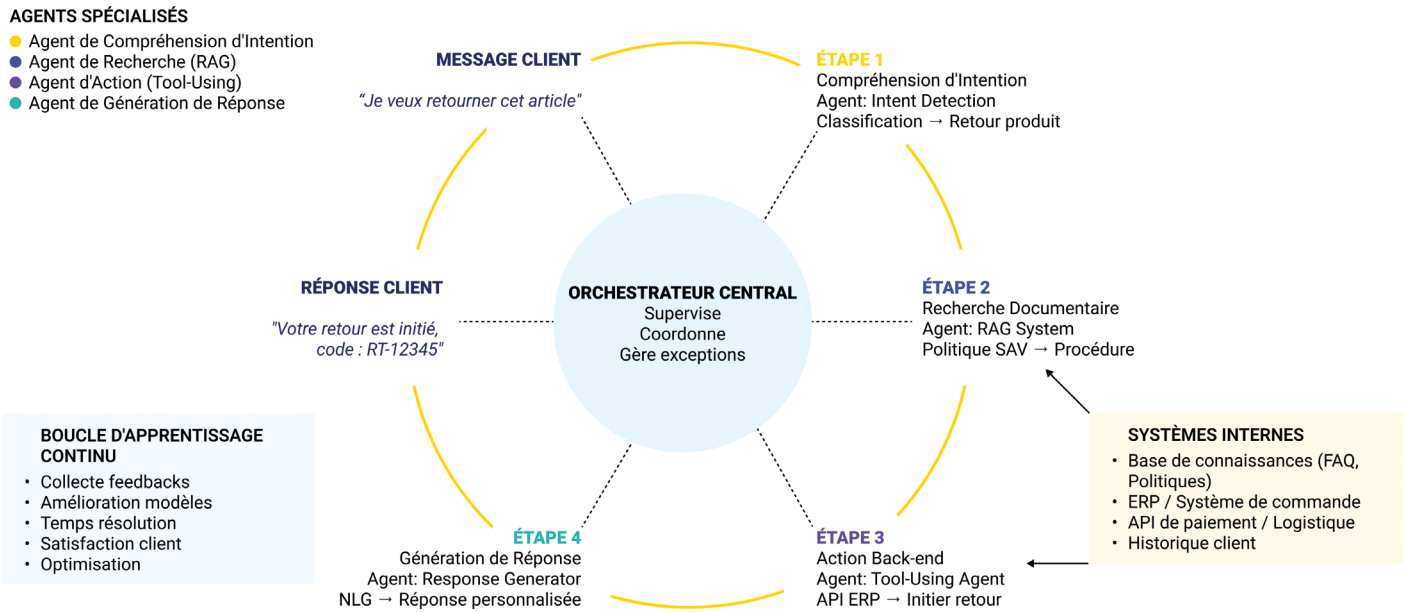
clients par mois, dans plusieurs langues et sur plusieurs zones géographiques. L'enjeu était de fournir une réponse rapide et pertinente 24h/24, en limitant la mobilisation du personnel sur des tâches répétitives.

La solution mise en œuvre par Klarna dépasse largement le cadre d'un simple *chatbot*. Derrière l'interface se cache un véritable système multiagents, pensé pour prendre en charge des tâches complexes, du premier message jusqu'à la résolution finale.

Le dispositif s'appuie sur plusieurs agents complémentaires, mobilisés en fonction du besoin réel exprimé par le client :

- Agent de classification d'intention : analyse les messages pour déterminer l'objectif réel du client (remboursement, suivi de commande, litige, etc.).
- Agent de recherche (RAG) : interroge la base documentaire interne pour identifier les bonnes politiques ou procédures.
- Agent d'action (*Tool-Using Agent*) : interagit de façon sécurisée avec les API internes de Klarna pour déclencher une opération concrète (comme effectuer un remboursement).
- Agent de génération de réponse : rédige une réponse synthétique, claire et empathique, que le client reçoit immédiatement.

Process Multi-Agents - Service Client Automatisé



Les résultats confirment la solidité du choix technologique et sa pertinence à grande échelle. À lui seul, le système absorbe un volume de travail équivalent à celui de plusieurs centaines de collaborateurs. Il gère en toute autonomie près des deux tiers des conversations clients, allégeant ainsi considérablement le travail des équipes.

Le temps moyen de résolution est passé de 11 minutes à moins de 2 minutes, tandis que le taux d'erreurs a diminué de 25 %, entraînant une nette amélioration de la satisfaction client. À l'échelle globale, Klarna prévoit un gain de rentabilité estimé à 40 millions de dollars grâce à cette automatisation orchestrée.

Source : https://www.zenml.io/llmops-database/ai-assistant-for-global-customer-service-automation

L'exemple de Klarna illustre donc une rupture à l'œuvre : on passe de l'optimisation d'étapes de la relation client à la réinvention d'un processus client complet. Le résultat est une relation de nouvelle génération : plus fluide, car elle élimine les ruptures ; et plus performante, car elle maximise les expertises là où elles comptent.

Cependant, le déploiement d'une telle vision à grande échelle présente de nombreux défis, comme nous allons le voir, en termes d'évolutivité et de souveraineté.

BÂTIR DES SOLUTIONS MULTIAGENTS SOUVERAINES ET ÉVOLUTIVES

La plupart des organisations a pris conscience des difficultés inhérentes au passage à l'échelle des projets informatiques mettant en jeu des grands volumes de données et de calculs, notamment liés à l'intelligence artificielle.

Mais les solutions mises en œuvre ont souvent consisté à externaliser pour éviter de gérer cette complexité en capitalisant sur les services managés des *hyperscalers* (Amazon Web Services, Google Cloud Platform, Microsoft Azure…).

Cependant cette approche est aujourd'hui remise en cause par le renouveau du protectionnisme commercial. Et si demain de nouvelles taxes multipliaient les factures par trois ou quatre ?

Garantir la souveraineté numérique

La souveraineté consiste à maintenir un contrôle juridique, opérationnel et technique sur les trois composantes essentielles de votre solution, sans dépendance envers des services tiers étrangers ou extra européens. L'objectif pour l'Europe est de forger un espace numérique fondé sur l'ouverture, la sécurité et la confiance.

Cette souveraineté s'articule autour de trois axes :
Le premier axe concerne la maîtrise des données. Il s'agit d'en garantir la confidentialité, de veiller à leur conformité stricte avec le cadre réglementaire européen (RGPD, *Data Act*), tout en les protégeant des législations extraterritoriales, telles que le *CLOUD Act* ou le *USA Freedom Act*.

Un point de vigilance majeur : la seule localisation des serveurs en Europe ne suffit pas à garantir la souveraineté si l'entité qui contrôle l'infrastructure reste soumise à une juridiction étrangère. D'où la nécessité d'une analyse approfondie des structures de gouvernance et des lois applicables.

Le second axe porte sur la transparence des modèles. Il s'agit de rendre leur fonctionnement intelligible, traçable et vérifiable, tout en veillant à leur alignement avec les valeurs de l'entreprise et à leur conformité aux réglementations en vigueur, en particulier l'*AI Act*.
Ce règlement européen, premier cadre juridique complet dédié à l'intelligence artificielle, classe les systèmes selon leur niveau de risque et vise à instaurer un écosystème d'IA éthique, responsable et digne de confiance.

Enfin, le troisième axe porte sur la conformité des infrastructures. Il s'agit de s'appuyer sur des environnements *cloud* alignés avec les cadres juridiques en vigueur, idéalement certifiés (tels que SecNumCloud ou HDS).

Investir dans des écosystèmes de données européens, à la fois sécurisés et capables d'interagir facilement entre eux, est ainsi devenu essentiel pour de nombreux acteurs.

Anticiper le passage à l'échelle (évolutivité)

La solution retenue doit être capable de passer du prototype à une mise en production à grande échelle, sans dégrader les performances ni entraîner de hausse significative des coûts.

Cette évolutivité maîtrisée se traduit par des bénéfices stratégiques concrets, tels que :

L'optimisation économique : l'évolutivité permet d'adapter automatiquement la puissance informatique selon les besoins réels, évitant ainsi de payer pour des ressources inutilisées.

Les services *cloud* fonctionnent en paiement à l'usage, ce qui évite les investissements lourds dès le départ.

Des outils comme Kubernetes – un système qui gère automatiquement le déploiement et l'utilisation des ressources – permettent de répartir intelligemment la charge de travail, pour une utilisation plus efficace et moins coûteuse.

Enfin, le choix de solutions *open source* peut réduire fortement les frais de licence à long terme.

L'amélioration de la résilience : Kubernetes renforce la fiabilité des systèmes grâce à ses capacités d'autoréparation : en cas de panne, il relance automatiquement les composants défaillants pour maintenir le service.

De leur côté, les systèmes multiagents, en répartissant les tâches entre plusieurs entités spécialisées, limitent les points de défaillance et renforcent la continuité de fonctionnement.

Cette robustesse structurelle devient un véritable avantage concurrentiel, en garantissant une disponibilité constante du service et une qualité maintenue, même en cas d'incident.

LE CADRE TECHNIQUE

Le succès d'une solution multiagents repose sur des choix techniques structurants. L'architecture s'articule autour de deux niveaux interdépendants : la couche applicative, qui constitue l'intelligence du système, et la couche d'hébergement, qui en forme la fondation souveraine et évolutive.

Le premier niveau établit les capacités de votre IA et ses modalités d'action, tandis que le second définit son environnement d'exécution. L'articulation de ces deux axes est cruciale pour concevoir un système alliant performance et maîtrise stratégique.

La couche applicative : choisir ses modèles et son orchestrateur

À ce niveau, l'enjeu technique se concentre sur l'architecture du système central. La sélection des modèles de langage (les « cerveaux ») et de l'orchestrateur (le « chef de projet ») constitue une décision déterminante qui influencera directement les performances, l'indépendance technologique et l'évolutivité de votre solution.

Le choix du LLM : une décision structurante

Dans une approche souveraine, deux options s'offrent à vous :

Les modèles propriétaires européens : cette approche offre la voie la plus directe vers la souveraineté. Ces modèles délivrent des performances de pointe avec la garantie d'un hébergement et d'un cadre contractuel alignés sur les exigences européennes.

Les modèles *open source* : ce terme signifie le plus souvent que les « poids » du modèle (soit les paramètres internes) sont rendus publics, autorisant une entreprise à les télécharger et à les opérer sur ses propres infrastructures. Le code source fourni permet de faire fonctionner le modèle dans sa phase dite d'inférence — c'est-à-dire le moment où l'IA produit des réponses à partir de nouvelles données (par exemple : répondre à une question, générer un texte, analyser une image).

En revanche, le code d'entraînement (la phase d'apprentissage du modèle) est rarement partagé.

Cette approche offre une maîtrise technique essentielle sur l'hébergement et l'opération du modèle, ce qui réduit l'effet de « boîte noire ». Même si la recette exacte de fabrication (les données d'entraînement, les étapes précises) reste confidentielle, l'entreprise garde le contrôle sur l'usage du modèle.

Il convient de souligner une distinction importante liée aux modèles *open source* non-européens (Llama 3, Qwen, etc.) : bien que leur origine puisse soulever des questions, leur caractère ouvert constitue une garantie technique fondamentale. Déployés sur une infrastructure souveraine, on s'assure qu'aucune donnée ne quitte l'entreprise et on garde un contrôle total sur l'exécution. C'est précisément ce qui les distingue des modèles propriétaires fermés, sur lesquels l'entreprise n'a ni visibilité, ni maîtrise complète.

Cependant, opter pour l'*open source* implique des responsabilités. Il faut anticiper trois points clés :

- Évaluer la transparence et les biais du modèle : l'absence de visibilité sur les données d'entraînement empêche de comprendre les biais inhérents du modèle (culturels, politiques, etc.). Il est donc indispensable de se renseigner sur les études d'alignement et, surtout, de ne faire confiance qu'à ses propres tests. La bonne pratique consiste à évaluer le comportement du modèle sur des *benchmarks* reconnus qui mesurent la toxicité, la véracité des informations ou les biais. Des outils comme ToxiGen, TruthfulQA ou DecodingTrust permettent par exemple de quantifier ces risques avant tout déploiement.

- Anticiper la maintenance : sans support commercial, il faut des compétences internes solides pour gérer les mises à jour, corriger les failles et assurer la pérennité de la solution.

- Vérifier les licences : le paysage des licences IA est complexe. Il est indispensable d'analyser la licence du modèle choisi pour s'assurer qu'elle est compatible avec les objectifs commerciaux de l'entreprise.

Bien entendu, le bon modèle ne se choisit pas théoriquement. Il faut le confronter au réel, l'évaluer sur vos propres cas d'usage, et voir comment il se comporte en situation. Le choix optimal nécessite une évaluation selon cinq critères :

- La performance : le modèle est-il réellement efficace sur vos langues cibles (notamment le français) et pour vos tâches spécifiques ?

- La fiabilité et l'alignement : ce critère est un prolongement direct des risques liés aux modèles *open source*. Le modèle présente-t-il des biais inacceptables ? Génère-t-il des contenus toxiques ? Peut-on se fier à la véracité de ses réponses ? L'évaluation sur des benchmarks (comme TruthfulQA) n'est plus une option, mais une nécessité.

- Le coût d'exploitation : quel sera son coût réel en production (inférence, maintenance) ?

- La licence : est-elle compatible avec votre usage commercial et vos objectifs ?

- La souveraineté : pouvez-vous l'héberger et le contrôler sur votre propre infrastructure ?

La bonne pratique : tester avant de choisir
Une erreur courante est de choisir un modèle uniquement sur sa réputation. La seule méthode fiable est de passer par une phase d'évaluation concrète pour trouver la meilleure combinaison modèle + cas d'usage.

Un exemple concret : la relation client
Pour des tâches très spécialisées comme le tri d'e-mails ou la catégorisation de demandes, il n'est pas toujours pertinent d'utiliser un grand modèle généraliste, qui peut s'avérer coûteux et surdimensionné. Dans ce cas, l'utilisation de petits modèles spécialisés (*Small Language Models*) est souvent un bien meilleur compromis. Ils sont plus légers, plus rapides et offrent un excellent rapport coût / performance pour des missions précises. C'est ce type d'arbitrage que seule une phase de test peut révéler.

Voici à la page suivante les principaux modèles *open source* disponibles au moment de la rédaction de cet article, avec leurs caractéristiques clés pour veiller à faire le bon arbitrage :

L'orchestrateur : le chef de projet de votre IA
Au cœur de votre système multiagents, l'orchestrateur agit comme un véritable coordinateur intelligent. Sa mission ? Analyser les demandes complexes, les décomposer en sous-tâches, confier chaque mission à l'agent le plus adapté, puis assembler les résultats partiels en une réponse cohérente et complète.

Deux approches technologiques s'offrent à vous : Les *frameworks open source* : ces outils offrent une flexibilité maximale pour construire des logiques adaptées à vos besoins métier. On trouve ici des outils de référence comme :

- LangGraph : idéal pour gérer des *workflows* complexes et cycliques entre agents, permettant des boucles de *feedback* et des processus décisionnels élaborés.

- CrewAI : se spécialise dans la collaboration entre agents aux rôles définis, facilitant la création d'équipes virtuelles pour accomplir des tâches.

Famille de modèles	Entreprise	Nationalité	Taille	Données d'entraînements	Code d'entraînement	Licence	Utilisation commerciale
Qwen 3	Alibaba	Chinoise	0.6B, 1.7B, 4B, 8B, 14B, 32B	Non	Non	Apache 2.0	Oui
DeepSeek R1	DeepSeek	Chinoise	1.5B, 7B, 8B, 14B, 32B, 70B	Non	Non	MIT/Apache 2.0	Oui
ERNIE 4.5	Baidu	Chinoise	0.3B, 21B, 28B	Non	Non	Apache 2.0	Oui
Magistral Small	Mistral AI	Française	24B	Non	Non	Apache 2.0	Oui
Mistral Small 3.2	Mistral AI	Française	24B	Non	Non	Apache 2.0	Oui
Gemma 3	Google	Américaine	1B, 4B, 12B, 27B	Non	Non	Gemma license	Oui
Phi 4	Microsoft	Américaine	3.8B, 5.6B, 14B	Non	Non	MIT	Oui
Llama 3	Meta	Américaine	1B, 3B, 8B, 70B	Non	Non	Llama 3 license	Oui
OLMo2	Allen Institute for AI	Américaine	1B, 7B, 13B, 32B	Oui	Oui	Apache 2.0	Oui
Molmo	Allen Institute for AI	Américaine	1B, 7B, 72B	Oui	Oui	Apache 2.0	Oui
Reka Flash 3	Reka AI	Américaine	21B	Non	Non	Apache 2.0	Oui
Llama Nemotron	Nvidia	Américaine	4.5B, 8B, 49B, 51B, 70B	Non	Non	NVIDIA Open Model License Agreement	Oui

- AutoGen (Microsoft) : permet de concevoir des écosystèmes d'agents conversationnels capables d'interagir et de résoudre des problèmes en équipe.

- Smolagents (Hugging Face) : se distingue par sa capacité à générer et exécuter du code en temps réel, offrant aux agents une grande flexibilité et un contrôle précis.

Les solutions logicielles intégrées : ces plateformes clé en main simplifient le déploiement, la supervision et la gouvernance des systèmes d'IA en entreprise.

- Dataiku : la plateforme propose une approche unifiée baptisée « LLM Mesh », conçue pour orchestrer plusieurs modèles de langage en parallèle. Elle permet ainsi de maintenir un contrôle rigoureux sur les coûts, les performances et la conformité réglementaire. Acteur français, Dataiku offre également l'avantage d'une proximité géographique et juridique, facilitant le respect des exigences souveraines.

- Synaptic Labs : développe des architectures sur mesure pour faciliter l'intégration d'agents d'IA dans les environnements d'entreprise existants, avec un focus particulier sur le monitoring et la gouvernance.

Le choix de votre orchestrateur doit s'appuyer sur trois piliers fondamentaux : sa capacité d'adaptation (modularité et extensibilité), sa transparence opérationnelle (*logs* détaillés, métriques de performance et de coût en temps réel), et sa

conformité aux standards de sécurité et de gouvernance qui régissent votre secteur d'activité (RGPD, SecNumCloud, etc.).

La couche d'hébergement : les acteurs européens et l'enjeu de la souveraineté

Le déploiement d'une solution d'IA nécessite une infrastructure robuste et évolutive. Au-delà des considérations techniques traditionnelles, ce choix revêt aujourd'hui une dimension stratégique majeure : celle de la souveraineté numérique.

L'utilisation de modèles d'IA implique une dépendance technologique dont il convient de mesurer les implications. La question évolue : au-delà de la performance pure, comment maintenir le contrôle sur ses données et préserver son autonomie décisionnelle ?

Les acteurs européens

Le marché européen du *cloud* se structure autour de deux philosophies distinctes, chacune répondant à des priorités stratégiques différentes.

Les acteurs européens « natifs » tels que OVHcloud, Scaleway, ou 3DS Outscale (filiale de Dassault Systèmes) offrent des garanties grâce à leurs certifications SecNumCloud et HDS.

Ces acteurs ont considérablement enrichi leur offre d'IA. L'époque où ils se limitaient à l'infrastructure brute est révolue : GPU H100 de dernière génération, plateformes d'entraînement de modèles, et services de déploiement managés font désormais partie de leur catalogue. Le partenariat stratégique entre Outscale et Mistral AI, qui propose un LLM souverain, illustre cette montée en puissance.

Les alliances pour un « *Cloud* de confiance » optent pour une approche de compromis. Les coentreprises comme Bleu (Orange/Capgemini/Microsoft) ou S3NS (Thales/Google) combinent la richesse fonctionnelle des géants américains avec une gouvernance et des opérations sous contrôle français.

Cette approche présente néanmoins des limites : l'obtention de la certification SecNumCloud peut entraîner des délais dans l'accès aux dernières innovations. Les organisations doivent donc arbitrer entre souveraineté maximale et accès immédiat aux technologies les plus avancées.

Les *hyperscalers* américains : des offres de « souveraineté » adaptées

Face à la demande croissante, les géants américains ont développé des offres spécifiques pour répondre aux exigences du marché européen. AWS propose son *European Sovereign Cloud* opéré depuis l'Allemagne par du personnel européen, Google déploie ses *Sovereign Solutions* avec différents niveaux de contrôle, et Microsoft développe son *Cloud for Sovereignty* avec un ensemble de garanties dédiées.

Malgré ces efforts techniques, ces solutions restent soumises au droit américain, notamment au *CLOUD Act*. Cette réalité constitue une différence fondamentale avec les offres européennes, qui seules peuvent garantir une réelle protection contre les lois extraterritoriales.

Ce point de divergence reste déterminant pour de nombreuses organisations, en fonction de leur secteur d'activité et de leurs exigences en matière de conformité.

Un point de vigilance : l'écosystème de services

L'écosystème de services managés offert par l'hébergeur constitue un facteur déterminant dans la rapidité de déploiement des solutions d'IA. Les *hyperscalers* et les « *Clouds* de confiance » qui s'appuient sur leur technologie excellent dans ce domaine grâce à leurs catalogues étoffés de services prêts à l'emploi : plateformes MLOps, bases de données vectorielles, outils de *monitoring* et d'orchestration…

À l'inverse, les acteurs 100 % européens, même s'ils enrichissent rapidement leurs offres, restent encore majoritairement axés sur la fourniture d'infrastructures brutes (IaaS). Le choix implique donc un vrai arbitrage : gagner en rapidité grâce aux services managés ou garder la main sur l'ensemble de l'infrastructure, avec en contrepartie une charge de développement et de maintenance plus lourde.

Comment choisir ?

Choisir une couche d'hébergement pour l'IA exige un arbitrage entre performance, latence, modèle économique et capacité à évoluer. S'y ajoute une dimension clé : garantir la souveraineté et l'accès à un écosystème de services solide.

Face à la complexité du choix, trois grandes approches stratégiques se dessinent, chacune répondant à des priorités différentes selon les organisations.
L'approche souveraine vise avant tout un contrôle total et une protection juridique maximale. Elle repose sur des infrastructures et des modèles européens certifiés (type SecNumCloud), ce qui implique généralement plus d'investissements en interne, que ce soit en expertise ou en développement de solutions sur mesure. C'est l'option privilégiée par les secteurs régulés ou manipulant des données particulièrement sensibles.

L'approche hybride, quant à elle, cherche un compromis entre souveraineté et efficacité. Elle peut s'appuyer sur des « *Clouds* de Confiance » — qui combinent la puissance technologique des *hyperscalers* américains avec une gouvernance locale — ou sur une architecture différenciée : les données sensibles sont hébergées de manière souveraine, tandis que les modèles sont utilisés à distance via API — c'est-à-dire une interface qui permet de se connecter à un service externe sans exposer les données ni l'infrastructure sous-jacente.

C'est une solution souvent choisie pour sa flexibilité et son pragmatisme.

Enfin, l'approche « *hyperscaler* maîtrisé » mise sur la rapidité d'exécution, en tirant parti des écosystèmes *cloud* mondiaux, tout en encadrant strictement les risques. Cela passe par des clauses contractuelles précises sur l'usage des données, l'obligation de localiser les *datacenters* en Europe, ou encore la mise en place d'audits réguliers. Elle nécessite une gouvernance solide, mais permet d'avancer vite.

Le bon choix dépendra toujours du contexte : la sensibilité des données, les contraintes réglementaires, les compétences disponibles en interne, ou encore le niveau de risque que l'organisation est prête à accepter. L'enjeu clé reste d'évaluer le degré de dépendance acceptable pour ses projets d'IA stratégiques.

CONCLUSION

FAIRE DE L'IA MULTIAGENTS UN LEVIER STRATÉGIQUE ET SOUVERAIN POUR LA RELATION CLIENT

L'émergence des systèmes multiagents marque une étape clé dans la manière dont les entreprises peuvent aujourd'hui réinventer leur relation client. En permettant à l'IA de ne plus seulement informer, mais aussi d'agir, ces systèmes transforment en profondeur l'expérience client qui devient plus fluide, plus cohérente, et surtout plus efficace. Ce n'est plus simplement une question de productivité, mais de création de valeur durable à chaque étape du parcours.

Les cas concrets présentés dans cet article l'ont démontré : qu'il s'agisse d'optimiser les ventes, d'alléger la charge des conseillers, ou d'anticiper les besoins des clients, les systèmes multiagents sont déjà à l'œuvre dans de nombreux secteurs. Ils ne sont plus un concept futuriste, mais bien une réalité opérationnelle. Pour autant, leur déploiement réussi repose sur trois piliers fondamentaux :

- Une architecture claire et évolutive, capable de s'adapter aux spécificités métiers.
- Un socle souverain et sûr, qui offre à l'entreprise un contrôle total sur ses données, ses modèles et ses infrastructures.
- Une mise à l'échelle maîtrisée, qui garantit la performance dans la durée, sans explosion des coûts ni compromis sur la qualité de service.

Dans un contexte qui conjugue des attentes élevées en matière de qualité de la relation client et des impératifs de compétitivité aigus, les systèmes intelligents peuvent devenir un levier stratégique significatif.

Contact :
Benjamin Hannache
Managing Partner
CGI Business Consulting
+33 (0)6 64 88 65 39