

Checklist : 10 points avant de mettre un agent IA en production

Un agent IA qui fonctionne en démo n'est pas prêt pour la production. Ces 10 points couvrent ce qui fait arrêter la plupart des agents après leur lancement : données exposées, réponses inventées, actions impossibles à vérifier, coûts qui dérivent.

Pour chaque point : pourquoi il compte, ce qu'il faut vérifier, le signal d'alerte, et un exemple. Les exemples sont fictifs.

Avant de construire

01. Cadrage du périmètre et des cas d'usage

Pourquoi. Un agent sans périmètre clair finit par répondre à tout, donc par se tromper sur tout.

- ☐ Le périmètre tient sur une page : ce que l'agent fait, ce qu'il ne fait pas, qui l'utilise.
- ☐ Un indicateur de succès est fixé avant le lancement : temps de traitement, taux de reprise humaine, coût par demande.
- ☐ Le processus métier existe et est documenté sans l'IA.

Signal d'alerte. « On verra ce que l'agent sait faire une fois en ligne. »

Exemple. Un agent « support » lancé sans liste de sujets autorisés se met à donner des conseils juridiques.

02. Permissions et contrôle des accès

Pourquoi. Un agent qui peut tout lire peut tout exposer.

- ☐ L'agent a son propre compte technique, jamais celui d'un administrateur ni le compte de service de l'application.
- ☐ Les droits sont appliqués par les systèmes (API, base de données), pas par une consigne dans le prompt.
- ☐ Chaque outil n'expose que les champs nécessaires, en lecture seule par défaut.

Signal d'alerte. « On a dit au modèle de ne pas regarder les autres dossiers. »

Exemple. Un agent RH qui interroge la paie avec un compte capable de lire tous les salaires : il suffit d'une question bien tournée pour qu'il en révèle un.

03. Sécurité des données et des secrets

Pourquoi. Les données envoyées au modèle et les clés d'accès sont les deux premières sources de fuite.

- ☐ Vous savez quelles données partent vers le modèle, où elles sont traitées et combien de temps elles sont conservées.
- ☐ Les données personnelles inutiles sont retirées ou masquées avant l'envoi.
- ☐ Les clés d'API et mots de passe sont dans un coffre-fort de secrets, jamais dans le code, le prompt ou les journaux.

Signal d'alerte. Une clé d'API visible dans un fichier de configuration ou un historique de conversation.

Exemple. Des journaux de débogage qui enregistrent les conversations complètes, numéros de contrat compris, et restent accessibles à toute l'équipe.

Pendant la conception

04. Garde-fous et maîtrise des actions

Pourquoi. Un agent qui agit (envoyer, modifier, rembourser) peut faire en une seconde ce qu'un humain mettrait une journée à réparer.

- ☐ La liste des actions autorisées est explicite ; tout le reste est bloqué techniquement par l'outil, pas seulement déconseillé dans le prompt.
- ☐ Les actions engageantes (paiement, remise, envoi externe, suppression) passent par une validation humaine.
- ☐ Quand l'information manque, l'agent le dit au lieu d'inventer. C'est testé avec des questions pièges et des demandes hors périmètre.

Signal d'alerte. L'agent dispose d'un outil « exécuter une requête » générique sur la base de production.

Exemple. Un agent commercial qui accorde une remise de 30 % parce qu'un prospect l'a demandé avec insistance.

05. Escalade humaine prévue

Pourquoi. Un agent qui ne sait pas passer la main bloque l'utilisateur, ou pire, improvise.

- ☐ Les cas à transmettre sont définis : sujet sensible, faible confiance, demande hors cadre, client mécontent.
- ☐ La transmission inclut tout le contexte : l'utilisateur n'a pas à tout répéter.

- ☐ Une personne est réellement disponible pour recevoir ces cas, avec un délai de réponse annoncé.

Signal d'alerte. « L'escalade envoie un mail à une boîte que personne ne lit. »

Exemple. Un agent RH transmet une contestation de salaire à la bonne personne, avec l'échange complet et le dossier concerné.

Pour exploiter

06. Observabilité et traçabilité

Pourquoi. Sans trace, impossible de savoir pourquoi l'agent a répondu ou agi ainsi, ni de le prouver.

- ☐ Chaque échange est journalisé : question, données consultées, outils appelés, réponse, décision.
- ☐ Vous pouvez retrouver en quelques minutes tout ce qui s'est passé pour une demande donnée.
- ☐ Les journaux sont protégés, avec une durée de conservation définie et sans secrets en clair.

Signal d'alerte. « Pour comprendre ce qu'il a fait, on relance la même question et on regarde. »

Exemple. Un client conteste une réponse : le journal montre la source consultée, qui était obsolète. Le problème est corrigé à la source en une heure.

07. Évaluation de la fiabilité

Pourquoi. « Ça a l'air de marcher » n'est pas une mesure.

- ☐ Un jeu de questions de test représentatif existe, avec les réponses attendues, cas difficiles compris.
- ☐ Il est rejoué à chaque changement de modèle, de prompt ou de source de données.
- ☐ Un seuil de qualité minimal est défini ; en dessous, pas de mise en production.

Signal d'alerte. Le modèle est mis à jour par le fournisseur et personne ne revérifie les réponses.

Exemple. Après une mise à jour du prompt, le taux de bonnes réponses sur le jeu de test passe de 92 % à 78 % : le changement est bloqué avant d'atteindre les utilisateurs.

08. Gestion des erreurs et rollback

Pourquoi. Les outils et les modèles tombent en panne : délais dépassés, quotas atteints, réponses vides.

- ☐ Chaque appel a un délai maximal, un nombre limité de nouvelles tentatives et un message de repli clair.
- ☐ L'agent peut être désactivé en un clic, sans bloquer le service : retour au processus humain.
- ☐ La version précédente (prompt, modèle, configuration) peut être restaurée rapidement.

Signal d'alerte. Pour arrêter l'agent, il faut redéployer toute l'application.

Exemple. Le fournisseur du modèle est indisponible pendant deux heures : les demandes basculent automatiquement vers l'équipe, avec un message d'attente.

09. Maîtrise des coûts Cloud et IA

Pourquoi. Le coût d'un agent ne se limite pas au modèle, et il peut dériver sans prévenir.

- ☐ Le coût par demande est mesuré : appels au modèle, infrastructure, outils, temps humain de reprise.
- ☐ Des alertes et des plafonds de budget sont en place.
- ☐ Le modèle est adapté à la difficulté : un petit modèle pour les demandes simples, un plus puissant seulement quand c'est nécessaire.

Signal d'alerte. La facture du mois est découverte en fin de mois.

Exemple. Une boucle d'appels entre deux outils multiplie le coût par dix en une nuit ; un plafond quotidien l'aurait arrêtée au bout d'une heure.

Le jour J

10. Préparation au lancement

Pourquoi. Un lancement mal préparé transforme un bon agent en mauvaise expérience pour les premiers utilisateurs, ceux qui comptent le plus.

- ☐ Le lancement est progressif : un groupe pilote, un canal ou un pourcentage du trafic, avant tout le monde.
- ☐ Les utilisateurs savent qu'ils parlent à une IA, et comment joindre une personne.
- ☐ Un responsable est désigné pour suivre les premiers jours : indicateurs, retours, incidents.

Signal d'alerte. Lancement le vendredi soir, pour tous les clients, sans personne d'astreinte.

Exemple. Une semaine de pilote sur 10 % des demandes révèle trois sujets mal couverts. Ils sont corrigés avant l'ouverture à tous.

Par où commencer

Un point non coché n'est pas forcément bloquant. Mais si vous ne deviez en traiter que trois, commencez par les points **02 (permissions)**, **04 (garde-fous)** et **06 (traçabilité)** : ce sont eux qui évitent les arrêts en production.

Vous voulez un regard extérieur sur votre projet ? Je propose un premier échange de 30 minutes, sans engagement : cloudatauditops.com.

Adil Laghdaif, architecte Cloud & IA — CloudAuditOps