
L'IA utile, sans perdre le contrôle.

Comprendre les principes. Choisir les modèles.
Automatiser ce qui compte.

PROMPT → MÉMOIRE → SYSTÈME D'INFORMATION

Depuis près de vingt ans, je développe des logiciels et accompagne des équipes. Aujourd'hui, l'IA prolonge ce travail : relier vos outils, rendre vos processus plus fluides et garder la main sur leur fonctionnement.

Jules Lepoivre · Fondateur de Saki Studio

Édition offerte · Version 1.1 · Mise à jour le 9 septembre 2026

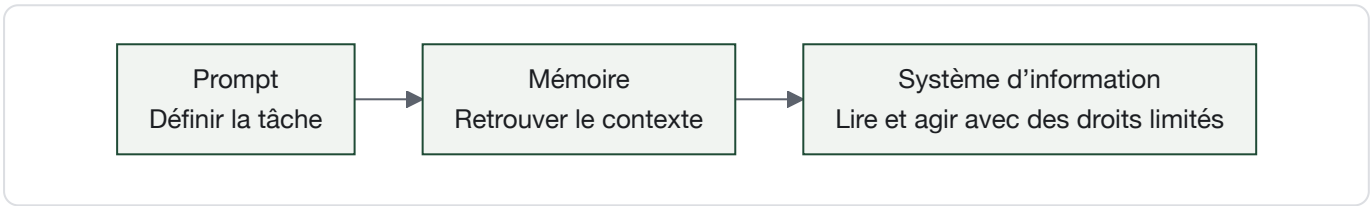
Sécurité, souveraineté, reproductibilité et explicabilité à chaque étape. Le **harness**, ou cadre d'exécution, organise leur mise en œuvre.

01 COMPRENDRE — Du prompt au système d'information

L'objectif : rendre un travail précis plus simple, plus fiable et plus facile à transmettre.

PROMPT	MÉMOIRE	SYSTÈME D'INFORMATION
Bien demander Une consigne et un résultat attendu.	Capitaliser Des règles et décisions réutilisables.	Agir au bon endroit Des données à jour et des outils reliés.
« Prépare cette réponse. »	« Applique notre façon de travailler. »	« Prépare le brouillon dans le bon dossier. »

Schéma 1 — Du briefing à l'action



Une progression selon le besoin : une tâche ponctuelle peut rester au premier niveau.

Le prompt : déléguer une tâche compréhensible

Précisez l'objectif, les informations utilisables, les limites et le format attendu. Une demande comme « fais-moi un devis » laisse trop de place à l'interprétation. Une demande qui fournit le besoin, le catalogue et les points à faire valider définit un travail contrôlable.

La mémoire : ne plus recommencer le briefing

Organisez les informations qui doivent survivre à une conversation : vocabulaire, décisions approuvées, contraintes et exemples. Le système doit retrouver le contexte pertinent, pas recopier toutes les archives dans chaque requête. Cette approche relève de la gestion du contexte, pas nécessairement de l'entraînement d'un nouveau modèle. [01]

Le système d'information : passer du texte au travail accompli

La messagerie, le logiciel clients, la facturation et les dossiers partagés constituent votre système d'information. Une automatisation utile y lit les bonnes données et y prépare une action autorisée. Les montants, statuts et droits d'accès restent sous le contrôle des logiciels métier.

Une progression, pas un classement. Un bon prompt peut suffire pour une tâche ponctuelle. La mémoire sert la continuité. L'intégration sert l'exécution. Sécurité, souveraineté, reproductibilité et explicabilité traversent les trois niveaux.

Repère : un modèle de langage produit des réponses plausibles, qui peuvent être fausses. Il ne constitue pas, à lui seul, une base de vérité pour votre entreprise. [03]

02 MIEUX TRAVAILLER — Un bon briefing, une mémoire utile

Commencez par une méthode que vous pouvez relire, corriger et transmettre à un collègue.

Un modèle de demande à réutiliser

Objectif : prépare [le livrable] pour [le destinataire]. **Contexte** : utilise uniquement [les documents autorisés].
Règles : respecte [le ton, les contraintes, les éléments interdits]. **Preuves** : indique les sources des faits importants et signale les contradictions. **Sortie** : fournis [le format], puis les points à vérifier. **Explication** : relie les faits aux sources, distingue les hypothèses et signale les informations manquantes. **En cas de manque** : écris « information manquante » ; n’invente pas. **Autorisation** : prépare un brouillon. N’envoie et ne modifie rien.

Demandez un résultat observable plutôt qu’un rôle prestigieux. Ajoutez un exemple de bonne réponse lorsque le format ou le ton compte. Le prompt aide à guider le modèle ; il ne remplace pas les contrôles techniques. [01, 07]

La mémoire minimale tient dans quatre documents

À CONSERVER	CONTENU UTILE
Votre activité	Offres, vocabulaire, publics et périmètre d’intervention.
Vos règles	Ce qui est autorisé, interdit ou soumis à validation.
Vos exemples	Quelques livrables approuvés, avec leurs sources.
Vos décisions	Choix validés, date, responsable et raison du choix.

Chaque information conservée doit avoir une source, une date, une version et un responsable. Séparez les clients et projets ; corrigez ou supprimez les éléments périmés. Une suggestion de l’IA ne devient pas automatiquement une règle approuvée. Pour chaque traitement, identifiez aussi les extraits de mémoire réellement fournis au modèle.

Ce que cela change concrètement

Sans mémoire, vous réexpliquez vos offres, votre ton et les décisions déjà prises. **Avec une mémoire sélectionnée**, la demande peut devenir : « Prépare la relance de ce dossier selon nos règles commerciales. » Le système récupère alors uniquement les éléments autorisés et pertinents.

Moins de texte saisi ne signifie pas forcément moins de tokens facturés. La mémoire transmise au modèle occupe aussi du contexte. Mesurez séparément le temps de briefing et le coût des appels. [17]

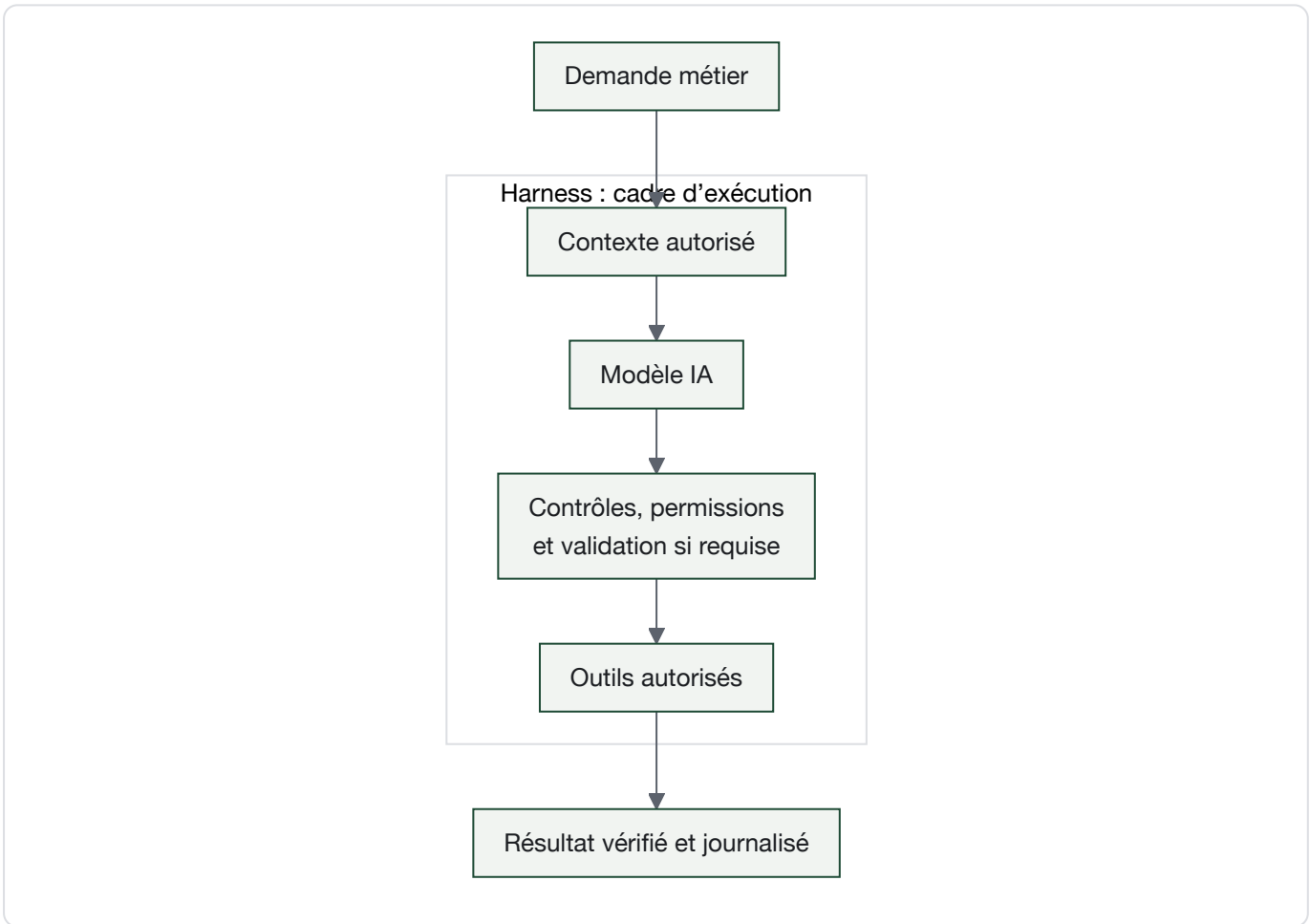
03 ENCADRER — Le harness, le cadre d'exécution de l'IA

Le **harness** désigne la couche logicielle qui prépare les entrées, appelle le modèle, orchestre les outils et gère l'état du travail. Il permet d'inscrire une demande dans un processus suivi. Les architectures d'agents décrites par OpenAI et Anthropic associent notamment gestion du contexte, exécution d'outils et continuité entre les étapes. [21, 22]

Pour une équipe, on peut simplement parler de **cadre d'exécution**. Voici les fonctions à prévoir dans un montage professionnel ; leur présence doit être vérifiée dans l'outil retenu.

FONCTION	QUESTION CONCRÈTE
Contexte et mémoire	Quels documents et quelles règles sont autorisés pour ce dossier ?
Outils et permissions	Que peut lire, préparer ou modifier le système, avec quels droits ?
Déroulement et limites	Quelle étape vient ensuite ? Quand arrêter, alerter ou reprendre ?
Contrôles et validations	Quels champs, sources et actions vérifier avant de continuer ?
Traces et versions	Comment retrouver le traitement, ses paramètres et son résultat réel ?

Schéma 2 — Un cadre entre la demande et les outils



Vue simplifiée : les appels d'outils passent par les contrôles prévus. Une réponse du modèle ne vaut pas autorisation d'exécuter une action.

Commencer avec un cadre léger

Pour préparer une réponse commerciale, un premier harness peut se limiter à une consigne versionnée, une fiche d'offres approuvée, un appel au modèle, une vérification des champs et un brouillon à relire. Les plafonds de coût, les droits et l'arrêt sur erreur sont configurés dans le logiciel.

Un outil de workflow existant ou un petit service sur mesure peut porter ce cadre. L'ensemble reste à tester : le mot « harness » ne garantit, à lui seul, ni la sécurité ni la qualité du résultat.

L'automatisation se construit autour du modèle : contexte utile, règles explicites, outils limités et résultats vérifiables.

04 FIABILISER — Reproduire un traitement et expliquer son résultat

Avant de confier une tâche récurrente à l'IA, posez deux questions : « **Pourra-t-on la rejouer dans des conditions connues ?** » et « **Pourra-t-on comprendre et vérifier ce qui a été fait ?** »

La reproductibilité : retrouver les conditions et mesurer la stabilité

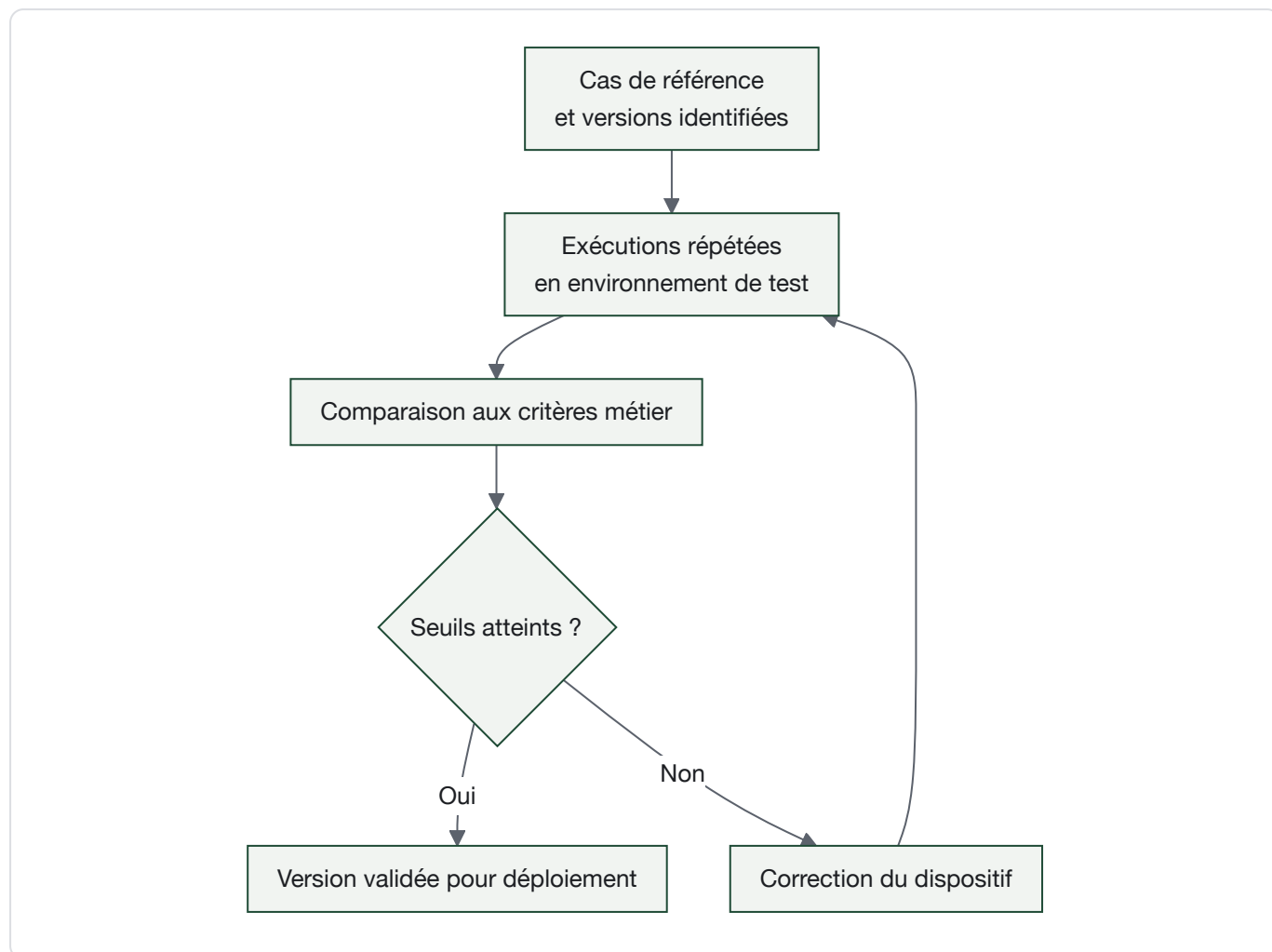
Les réponses d'un modèle peuvent varier entre deux exécutions. Une évaluation sérieuse répète donc les essais et examine le résultat obtenu dans les outils, au-delà du texte qui l'annonce. [23]

Pour votre pilote, distinguez **le processus rejouable**, **la stabilité du résultat métier** et **l'identité exacte du texte**. Une relance peut changer de formulation tout en respectant les mêmes faits. Un montant, un destinataire ou un droit d'accès exige des contrôles explicites.

À IDENTIFIER POUR CHAQUE TRAITEMENT	POURQUOI ?
Version du prompt, du workflow et du harness	Retrouver les consignes et le déroulement utilisés.
Modèle, version disponible, fournisseur effectif et réglages	Repérer un changement de modèle ou de route de secours.
Documents, extraits de mémoire et réponses d'outils utilisés	Reconstituer le contexte disponible au moment de l'appel.
Date de référence, identifiant et résultat réel	Vérifier les échéances, les reprises et l'absence de doublons.

Méthode proposée : préparez un lot de cas de référence, définissez les critères d'acceptation, puis rejouez-le après tout changement de prompt, de mémoire, de modèle ou de connecteur. Gardez quelques cas à part pour vérifier que les corrections fonctionnent aussi sur des exemples non utilisés pour les ajustements.

Schéma 3 — Tester avant chaque évolution



Exemple de critères : les champs obligatoires sont présents, les faits sont sourcés, aucune action interdite n'est exécutée et les cas ambigus sont transmis à un humain.

Pour un rejeu technique, utilisez des données de test et, si nécessaire, des réponses d'outils enregistrées ou simulées. Les envois, paiements et modifications réels doivent être désactivés. Pour une nouvelle exécution en production, relisez au contraire les données à jour et vérifiez les doublons.

Un traitement reproductible reste soumis à une vérification de justesse. Une erreur répétée de façon stable reste une erreur. Toute limite de rejeu — ancienne version du modèle indisponible, document supprimé — doit être indiquée.

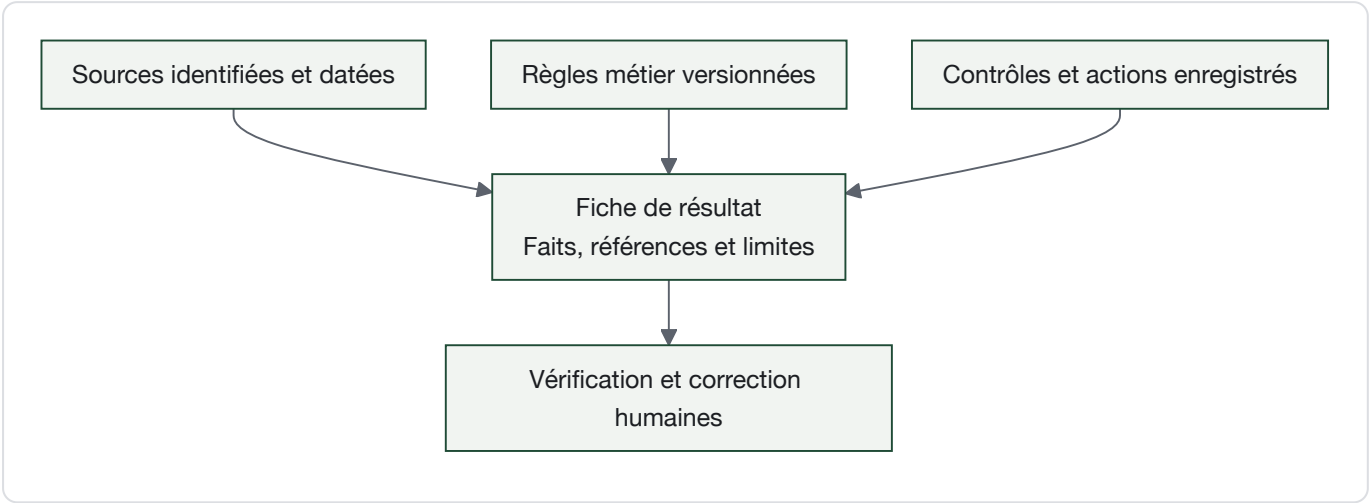
L'explicabilité : donner des raisons vérifiables

Au niveau métier, l'explicabilité consiste à rendre compréhensibles les faits retenus, les règles appliquées, les limites et les actions réalisées. Elle s'appuie sur la traçabilité : sources, contrôles et événements réellement enregistrés.

Une justification rédigée par le modèle peut omettre des éléments qui ont influencé sa réponse. Les recherches sur la fidélité des raisonnements affichés invitent à ne pas les considérer comme une preuve complète du fonctionnement interne. [24]

Format proposé pour accompagner une suggestion : résultat, références des faits utilisés, règle métier applicable, informations manquantes, vérifications effectuées et action restant à valider. Les contrôles logiciels doivent fournir leur propre statut ; le modèle ne s’attribue pas lui-même la validation.

Schéma 4 — Relier le résultat aux preuves



Le lecteur peut retrouver les éléments qui soutiennent le résultat, identifier une limite et demander une correction.

Exemple fictif : pourquoi proposer cette relance ?

ÉLÉMENT DE LA FICHE	CONTENU
Proposition	Préparer une relance du devis D-042.
Faits	Devis émis le 25 août 2026 ; statut « envoyé » au 9 septembre ; aucune réponse enregistrée dans le suivi.
Règle	Relance à partir de J+14 pour un devis toujours « envoyé » — règle commerciale v3.
Contrôles	Statut relu ; aucune relance déjà enregistrée pour cette échéance.
Limite	Un échange téléphonique non saisi dans le logiciel peut manquer.
Action	Brouillon préparé ; relecture requise ; aucun envoi automatique.

Cette fiche facilite la vérification du cas concret. Une référence doit permettre de retrouver le passage qui soutient le fait annoncé, avec les droits d’accès appropriés. La traçabilité du processus ne suffit pas à décrire intégralement les mécanismes internes du modèle.

Conservez des traces proportionnées. Privilégiez les références vers des sources versionnées et les métadonnées utiles. Les copies d’entrées, mémoires, sorties et journaux doivent avoir une finalité, des accès limités et une durée définie. Les obligations de protection des données continuent de s’appliquer aux traces d’audit. [25]

05 PROTÉGER — La sécurité avant l'autonomie

Plus un système peut lire, écrire ou envoyer, plus ses erreurs et ses accès doivent être encadrés.

1. Définir ce qui peut sortir

Classez les données avant de connecter un outil : publiques, internes, confidentielles ou particulièrement sensibles. Pour un premier essai, privilégiez des exemples fictifs ou réellement anonymisés. Retirer les noms ne suffit pas toujours à anonymiser un dossier. Faites valider les cas sensibles avec les personnes responsables de la sécurité et des données. [03, 04]

2. Donner uniquement les droits nécessaires

Utilisez des comptes professionnels et des accès dédiés, avec authentification renforcée. Un assistant de rédaction n'a pas besoin du droit d'envoyer, de payer ou de supprimer. Les autorisations doivent être contrôlées par le logiciel, pas seulement demandées dans le prompt. [07]

3. Traiter les documents comme des données, pas comme des ordres

Un e-mail ou un document peut contenir des instructions malveillantes destinées à détourner l'IA : c'est l'injection de prompt. Le contenu lu ne doit jamais pouvoir élargir les droits, choisir un destinataire externe ou déclencher une action interdite. La recherche documentaire ne supprime pas ce risque. [06]

4. Faire valider ce qui engage l'entreprise

Au démarrage, imposez une validation avant un envoi externe, un engagement commercial, une modification de coordonnées bancaires ou une suppression. Présentez au validateur le destinataire, les données sources et l'action exacte, pas seulement un bouton « OK ». [07]

5. Prévoir l'arrêt et la reprise

Fixez un budget, un nombre maximal de tentatives et un responsable des alertes. Conservez un journal proportionné : qui a demandé quoi, quelle version a été utilisée et quelle action a été exécutée. Prévenez les doublons et prévoyez une reprise manuelle.

6. Entretenir le dispositif

Protégez les clés d'accès, limitez la conservation des journaux, mettez à jour les composants et testez la restauration des données. L'IA ne dispense pas des bonnes pratiques de développement et d'exploitation. [04]

La CNIL : traduire la protection des données en décisions concrètes

Dès qu'un processus utilise des données personnelles — y compris dans les prompts, la mémoire ou les journaux — son cadrage doit intégrer le RGPD. Les recommandations de la **CNIL, Commission nationale de l'informatique et des libertés**, aident à définir les usages, les responsabilités et les précautions nécessaires. [03]

Les questions ci-dessous constituent une fiche de préparation, à adapter au traitement concerné. [25, 26]

QUESTION À TRANCHER	ÉLÉMENT À PRÉPARER
Pourquoi traiter ces données ?	Finalité précise et base légale adaptée, à documenter avant le traitement.
Quelles données sont nécessaires ?	Liste des champs utiles, exclusions et limitation de la mémoire conservée.
Qui intervient ?	Responsable de traitement, prestataires, rôles et contrats de sous-traitance lorsque ce rôle s'applique.
Où passent les données ?	Stockage, calcul, journaux, support, sous-traitants ultérieurs et transferts hors UE éventuels.
Comment informer et respecter les droits ?	Information compréhensible, point de contact et procédure d'accès, de rectification ou d'effacement selon les droits applicables.
Combien de temps conserver ?	Durées adaptées aux documents, extraits de mémoire et traces, avec procédure de suppression.

Le dossier de cadrage alimente le registre des traitements lorsque celui-ci est requis. Les accès, contrats, formations et contrôles doivent ensuite être maintenus dans la durée. Pour l'information des personnes, distinguez clairement les usages de leurs données ; une fiche destinée à l'entraînement d'un modèle ne décrit pas automatiquement votre automatisation métier. [26, 29]

AIPD : vérifier le besoin avant le démarrage. Une analyse d'impact relative à la protection des données est obligatoire lorsque le traitement est susceptible d'engendrer un risque élevé pour les droits et libertés des personnes. Elle n'est pas systématique pour toute utilisation d'IA. Si un risque résiduel élevé subsiste malgré les mesures prévues, une consultation préalable de la CNIL est requise dans les conditions applicables. [27]

Décisions importantes : prévoir un examen spécifique. L'article 22 du RGPD encadre les décisions fondées exclusivement sur un traitement automatisé qui produisent des effets juridiques ou affectent une personne de manière significative. Elles sont interdites en principe, avec des exceptions encadrées. Une validation humaine doit permettre un examen réel et une remise en cause du résultat. [28]

À faire valider avec le délégué à la protection des données, lorsqu'il existe, ou un conseil compétent : finalité, base légale, transferts, besoin d'AIPD et cas de décisions automatisées. Ces repères facilitent le cadrage ; ils ne constituent ni un avis juridique personnalisé ni une certification de conformité.

Point de départ recommandé : lecture → proposition → validation humaine. L'autonomie vient ensuite, uniquement pour des actions bornées, testées et réversibles.

06

GARDER LA MAIN — La souveraineté, au-delà du mot « local »

Gardez la maîtrise de vos données, de vos règles et de votre capacité à changer de fournisseur.

Quatre questions à poser avant de choisir

QUESTION	CE QU'IL FAUT VÉRIFIER
Où vont les données ?	Stockage, calcul du modèle, sauvegardes, journaux et accès de support.
Qui peut les utiliser ?	Fournisseur du modèle, hébergeur, intermédiaires et autres outils connectés.
Combien de temps ?	Conservation, entraînement éventuel, effacement et conditions contractuelles.
Peut-on partir ?	Export des données, règles et workflows ; procédure de migration testable.

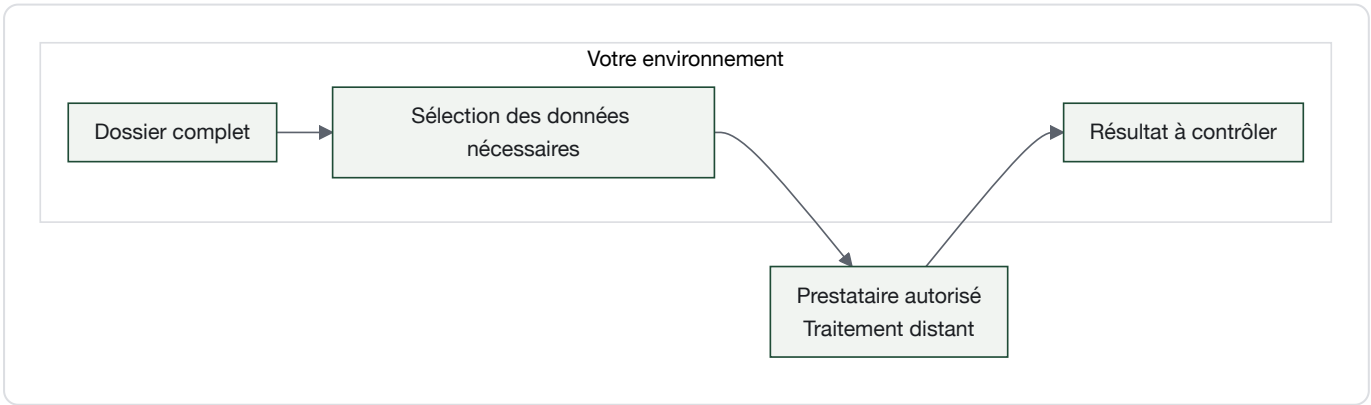
Un hébergement dans l’Union européenne n’est pas, à lui seul, une garantie de souveraineté ou de conformité. Examinez aussi les opérateurs, leurs accès et les transferts éventuels. Pour des données personnelles, les responsabilités et les garanties contractuelles doivent être encadrées. [03]

Trois architectures à examiner selon le risque

APPROCHE	COMPROMIS À EXAMINER
Service cloud encadré	Mise en route facilitée ; dépendance au fournisseur et règles de données à vérifier.
Approche hybride	Données conservées chez vous ; seuls des extraits autorisés sont envoyés au modèle distant.
Traitement local ou privé	Davantage de maîtrise technique ; capacités, coût matériel, maintenance et sécurité à assumer.

Mémoire locale ≠ traitement local. Si un agent transmet un extrait à un modèle distant, cet extrait quitte votre environnement. Orchbun distingue précisément le stockage local de la mémoire et les conditions du fournisseur utilisé par l’agent. [17]

Schéma 5 — Voir ce qui quitte votre environnement



Exemple d’architecture hybride : les données sélectionnées quittent l’environnement. Vérifiez aussi les journaux, la conservation et les accès chez chaque intervenant. Réduire les données transmises ne garantit pas leur anonymat. [03, 14, 15]

Modèle téléchargeable ≠ souveraineté automatique. Pouvoir héberger un modèle est une possibilité d’architecture ; cela ne dit rien du lieu où une requête API donnée est réellement traitée.

OpenRouter : une référence tarifaire, pas une garantie automatique. Vérifiez les fournisseurs autorisés, la rétention et les routes de secours. « Pas d’entraînement » ne veut pas dire « aucune conservation ». Le réglage zéro rétention ne couvre pas nécessairement les outils tiers, comme la recherche web. [14, 15]

Les mémoires persistantes et les connexions entre services augmentent les données en circulation : la CNIL souligne ce point dans sa note sur l’IA agentique de juillet 2026. [05]

07 CHOISIR — Modèles : prix, performance et fiabilité

Photographie au 9 septembre 2026. Quatre modèles de pointe pour présélectionner, pas un classement universel.

ÉDITEUR / MODÈLE	ENTRÉE \$/1 M	SORTIE \$/1 M	INDICE AA*
Anthropic Claude Fable 5.1 [08]	10,00	50,00	53
OpenAI GPT-6 Astra [09]	10,00	50,00	53
Moonshot AI Kimi K3 [10]	3,00	15,00	44
MiniMax MiniMax M3 [11]	0,30	1,20	30

Un token est une petite unité de texte. L'entrée comprend la consigne et le contexte envoyés ; la sortie comprend la réponse et le raisonnement facturés. API désigne ici l'accès au modèle depuis un logiciel. [18]

Tarifs texte standard des fournisseurs éditeurs affichés sur OpenRouter, sans cache, batch, Flex, mode rapide ni majoration de contexte long. Dollars américains, hors frais de plateforme et taxes. Il s'agit des prix API, pas des abonnements aux applications de chat. [08–11]

D'autres fournisseurs affichent des planchers plus bas : Kimi K3 à 2,40 \$ / 12 \$; MiniMax M3 à 0,23 \$ / 0,96 \$. Ces routes ne garantissent pas la même disponibilité, la même latence ou les mêmes règles de données. [10, 11]

Comment lire la « fiabilité »

L'indice AA n'est pas un pourcentage de réponses correctes. Il synthétise des capacités évaluées par Artificial Analysis. Les scores arrondis ci-dessus correspondent à Fable 5.1 « max with fallback », GPT-6 Astra « max », Kimi K3 « max » et MiniMax-M3. Les réglages, notamment le secours de Fable, font partie du résultat. [12]

La fiabilité dans votre activité se mesure autrement : exactitude des données, respect des consignes, bon usage des outils, gestion des cas inconnus et stabilité de l'exécution. Les tests publics servent à choisir les candidats au pilote ; ils ne prouvent pas leur fiabilité sur vos dossiers. Pour une automatisation, évaluez le couple **modèle + harness**, avec son contexte, ses outils et ses contrôles : changer le cadre d'exécution peut changer les résultats. [23]

Ma présélection pragmatique

Fable 5.1 et GPT-6 Astra : commencer par eux pour établir une référence sur les cas difficiles. **Kimi K3** : tester comme compromis entre capacité et coût. **MiniMax M3** : tester en priorité sur les tâches bornées et répétitives. Ce sont des orientations de test, pas des garanties de qualité.

Vigilance confidentialité : la fiche OpenRouter de Fable 5.1 avertit d'une conservation des données et de l'absence de zéro rétention selon la politique Anthropic indiquée. Ne présumez pas qu'il convient à un besoin strict de non-conservation. [08]

08 MESURER — Le vrai coût : un dossier correctement traité

Un modèle moins cher n'est intéressant que s'il produit le résultat attendu avec un coût de contrôle acceptable.

Un exemple de coût API, pas un devis de projet

Hypothèse illustrative : 1 000 traitements, chacun consommant 4 000 tokens d'entrée et 1 000 tokens de sortie facturés au total. Cela représente 4 millions de tokens d'entrée et 1 million de sortie.

MODÈLE	CALCUL	COÛT API
Claude Fable 5.1	$4 \times 10 \$ + 1 \times 50 \$$	90 \$
GPT-6 Astra	$4 \times 10 \$ + 1 \times 50 \$$	90 \$
Kimi K3	$4 \times 3 \$ + 1 \times 15 \$$	27 \$
MiniMax M3	$4 \times 0,30 \$ + 1 \times 1,20 \$$	2,40 \$

Calculs issus des tarifs de la page précédente. Le volume réel varie selon le modèle. Les tokens de raisonnement sont facturés comme sortie : une réponse courte peut donc coûter davantage que son texte visible. OpenRouter affiche des frais de plateforme de 5,5 % à l'achat de crédits ; ils sont exclus ici, ainsi que les taxes, outils et hébergement. [13, 18]

Le test métier à faire avant de déployer

Préparez **30 exemples anonymisés** : 20 cas courants, 5 cas incomplets ou ambigus, 5 cas de risque ou d'erreur. Définissez le résultat acceptable avant de comparer. Utilisez les mêmes documents, règles et outils pour chaque candidat ; notez le modèle, le fournisseur et les réglages.

Mesurez le résultat correct du premier coup, les inventions, les actions interdites, la bonne escalade vers un humain, les minutes de correction, la latence et le coût complet. Répétez les cas sensibles. Un premier lot réussi ne prouve pas l'absence future d'erreurs. Ajoutez la stabilité entre répétitions et la proportion de résultats dont les sources, les règles et les actions sont vérifiables. Conservez toutes les tentatives du test, pas seulement la meilleure réponse.

Coût par dossier accepté = (coût des appels, outils et reprises + coût du temps humain de contrôle et correction) ÷ nombre de dossiers acceptés. Ajoutez la maintenance pour juger le coût d'exploitation.

Mesurer le gain sans le survendre

Exemple fictif : 200 tâches par mois passent de 8 à 3 minutes, contrôle compris. Le gain est de **16 h 40 par mois**, avant maintenance. C'est de la capacité libérée, pas automatiquement une économie de trésorerie. Déduisez aussi l'intégration et la formation pour juger la rentabilité du projet.

Choisissez le plus petit dispositif capable de produire un résultat utile et vérifiable.

1. Automatiser sans IA quand la règle est claire

Déplacer un fichier, notifier une échéance, recopier un champ ou calculer un montant ne demande pas nécessairement un modèle de langage. Commencez par les fonctions déjà disponibles dans vos logiciels. Réservez l'IA à ce qui exige de comprendre un texte ou d'en produire un.

2. Ajouter une seule étape IA à un processus classique

Un déclencheur récupère le contenu ; l'IA propose une classification ou un brouillon ; le logiciel vérifie le format ; un humain valide. Le chemin reste prédéfini. Ce modèle est plus facile à examiner qu'un système qui choisit librement toutes ses actions. La distinction entre workflows définis et agents autonomes est utile pour choisir cette complexité. [02]

3. Réutiliser une petite base de contexte approuvée

Pour les tâches récurrentes, fournissez la bonne procédure, le bon catalogue ou l'historique pertinent. Ne commencez pas par aspirer tous les dossiers de l'entreprise. Prévoyez qui maintient les documents, leur version et les droits d'accès.

4. Ne connecter l'action qu'après la preuve d'utilité

Démarrez avec une proposition. Ajoutez ensuite la création d'un brouillon ou d'une tâche interne. L'automatisation sans validation peut être réservée plus tard aux actions peu risquées, réversibles et dont les exceptions sont bien connues.

BESOIN	POINT DE DÉPART PROPOSÉ
Un usage individuel	Consigne réutilisable + documents autorisés + relecture.
Un enchaînement entre logiciels	Fonction native, puis workflow visuel dans n8n ou Make si nécessaire. [16, 19]
Des règles métier particulières	Petit connecteur ou service sur mesure, avec contrôles et journal.
Une décision variable à plusieurs étapes	Agent limité à des outils précis, seulement après évaluation.

Le modèle propose ; le logiciel vérifie ; l'humain arbitre ce qui engage. Le choix d'un outil ne remplace pas la définition du processus, de ses exceptions et de son responsable.

10 TROUVER UN PREMIER CAS — Trois usages pour les échanges quotidiens

Exemples de conception, non résultats clients. Le point de départ commun : préparer un brouillon plutôt qu’agir à votre place.

01 — Trier les demandes et préparer les réponses

Déclencheur : une nouvelle demande dans une boîte dédiée. **Contexte** : catégories de demandes, offres et exemples approuvés. **Résultat** : catégorie proposée, résumé, informations manquantes et brouillon de réponse.

Version de départ : traiter uniquement les messages sélectionnés manuellement. **Garde-fou** : aucune promesse de prix, de délai ou de disponibilité sans donnée source ; aucun envoi automatique. **Mesure** : temps de préparation, bon classement et proportion de brouillons réellement utilisables.

02 — Transformer des notes de réunion en actions

Déclencheur : dépôt de notes approuvées. **Contexte** : format de compte rendu et liste des projets. **Résultat** : décisions explicites, actions proposées, responsable et échéance lorsqu’ils ont été mentionnés.

Version de départ : produire un document, sans créer de tâches dans les outils. **Garde-fou** : ne pas transformer une idée discutée en décision ; laisser « à confirmer » pour un responsable ou une date non définis. **Mesure** : temps de finalisation et nombre de corrections. Préférez les notes écrites pour démarrer ; un enregistrement audio nécessite un cadre spécifique.

03 — Préparer une réponse de support avec ses sources

Déclencheur : un ticket entrant. **Contexte** : documentation à jour et procédures accessibles à l’utilisateur. **Résultat** : réponse proposée, liens vers les passages utiles et motif d’escalade éventuel.

Version de départ : suggestion visible seulement par l’équipe support. **Garde-fou** : ne pas inventer une fonctionnalité, une garantie ou une solution absente de la documentation. **Mesure** : temps de résolution, corrections et escalades justifiées. Maintenez une liste des questions auxquelles le système ne doit pas répondre seul.

Choisissez un flux étroit. « Les demandes de devis reçues sur cette boîte » est un meilleur premier périmètre que « toute la relation client ».

11 TROUVER UN PREMIER CAS — Trois usages pour l'administratif et le suivi

L'IA prépare et explique. Les règles, les calculs et les validations restent explicites.

04 — Extraire les informations d'un document

Déclencheur : ajout d'une facture, d'un bon de commande ou d'une fiche fournisseur. **Résultat** : une ligne dans une table de préparation, avec la référence du document et les champs extraits.

Version de départ : un type de document et quelques champs. **Garde-fou** : contrôler les totaux par calcul, détecter les doublons et les champs absents, conserver le lien vers la pièce. Aucun paiement ni changement de coordonnées bancaires automatique. **Mesure** : exactitude des champs et temps de correction. Pour un PDF natif, privilégiez d'abord son texte extractible.

05 — Préparer les relances de devis

Déclencheur : une échéance calculée à partir du statut réel du devis. **Contexte** : dernier échange, règles commerciales et informations de suivi. **Résultat** : rappel interne ou brouillon personnalisé.

Version de départ : une liste de devis à vérifier ; un modèle de message peut suffire sans IA. **Garde-fou** : relire le statut avant envoi, exclure les dossiers clos ou déjà relancés, ne jamais inventer une remise. **Mesure** : couverture des dossiers à suivre, doublons évités et temps de préparation.

06 — Produire une synthèse hebdomadaire sourcée

Déclencheur : une fréquence fixe. **Contexte** : indicateurs calculés par les outils de gestion et liste des événements de la période. **Résultat** : faits marquants, écarts observés, points de vigilance et liens vers les données.

Version de départ : une équipe, trois indicateurs, un destinataire interne. **Garde-fou** : distinguer les faits des hypothèses ; ne pas inventer la cause d'une variation. L'IA commente les chiffres, elle ne remplace pas les formules. **Mesure** : temps de préparation, traçabilité et utilité de la synthèse pour les destinataires.

La bonne question : « Quelle partie répétitive peut être préparée automatiquement ? » Pas : « Comment remplacer tout le processus par une IA ? »

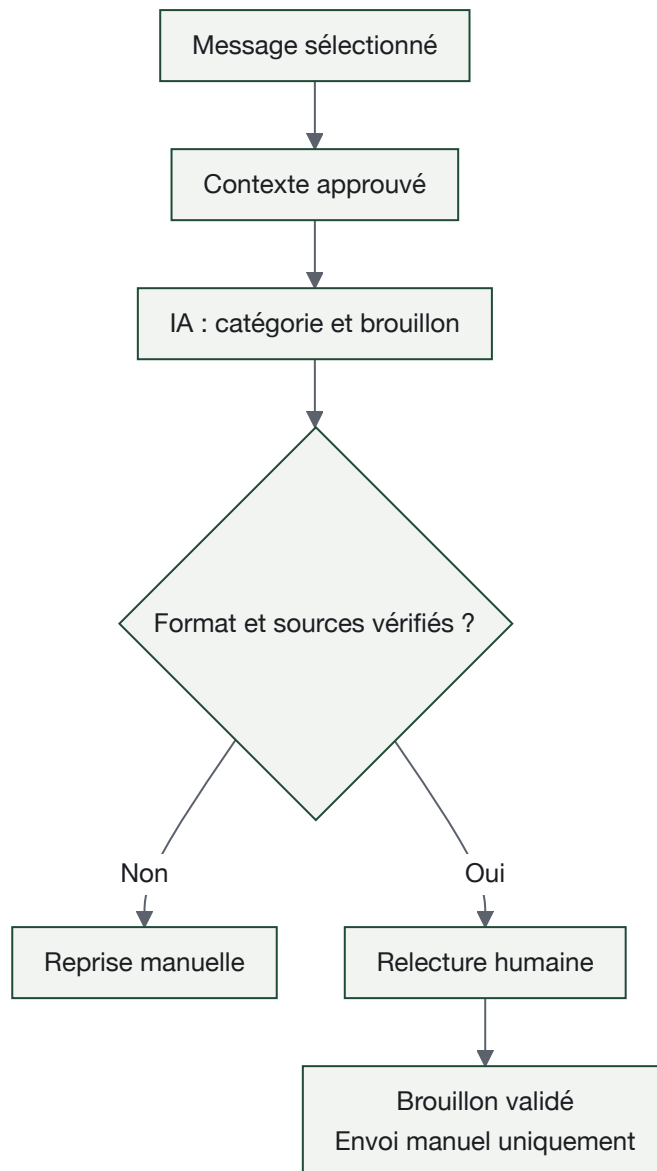
12

PASSER À L'ACTION — Votre premier workflow, en mode brouillon

Cas retenu : qualifier une demande commerciale et préparer la réponse. Un périmètre restreint et aucun envoi automatique.

ÉTAPE	MISE EN ŒUVRE PROPOSÉE
1. Sélectionner	Commencez avec dix messages fictifs ou anonymisés. Ensuite, utilisez une boîte dédiée ou un accès techniquement limité.
2. Fournir le contexte	Ajoutez une fiche d'offres à jour, les catégories et deux réponses approuvées. Ne transmettez pas tout l'historique client.
3. Faire proposer	L'IA renvoie : catégorie, résumé, informations manquantes et brouillon. Chaque fait commercial doit être relié à une source fournie.
4. Contrôler	Le workflow vérifie les champs, les catégories autorisées et les références. Les absences ou incohérences passent en file de revue.
5. Faire relire	Présentez le message original, le contexte, le brouillon et ses alertes. L'envoi reste une action humaine dans la messagerie.
6. Tracer	Enregistrez l'identifiant, les versions, les références du contexte, les contrôles, le statut, le coût et la validation humaine, avec accès et conservation limités.

Schéma 6 — Une première automatisation, avec relecture



La vérification automatique prépare la relecture. L'envoi reste une action de la personne responsable.

La consigne de l'étape IA

Classe cette demande dans « devis », « support » ou « autre ». Résume le besoin en trois phrases. Liste les informations manquantes. Propose une réponse de moins de 150 mots en utilisant uniquement les offres et règles fournies. N'invente aucun tarif, délai ou engagement. Si une information est absente ou contradictoire, indique-le pour la relecture.

Les réglages que le prompt ne peut pas remplacer

Aucun outil d'envoi, de paiement ou de suppression accessible à l'agent. Un identifiant de traitement pour éviter les doublons. Un plafond de coût, un nombre limité de tentatives et une alerte en cas d'échec. Les destinataires et les accès ne sont jamais déduits d'une instruction contenue dans l'e-mail. [06, 07]

Avant toute évolution : rejouez les cas de référence en mode test, sans envoi. Vérifiez aussi un tarif absent, une source contradictoire, un message déjà traité et une instruction malveillante dans le contenu. Le validateur doit pouvoir retrouver les raisons de la suggestion.

Pour le montage : ce schéma peut être réalisé dans un outil de workflow ou un petit service sur mesure. n8n documente notamment la validation humaine des appels d'outils ; vérifiez le périmètre réel des accès accordés par chaque connecteur. [16]

Critère de passage à l'étape suivante : les utilisateurs gagnent du temps, comprennent les alertes et savent reprendre la main. Une démonstration impressionnante ne suffit pas.

13 CADRER LE PILOTE — Une tâche. Un responsable. Un résultat.

Remplissez cette fiche avant d'acheter un nouvel abonnement ou de connecter un outil à vos données.

À DÉFINIR	VOTRE RÉPONSE
Tâche répétitive et résultat attendu	
Volume mensuel et temps actuel	
Données nécessaires et accès autorisés	
Cadre RGPD, intervenants et durées de conservation	
Action interdite ou soumise à validation	
Responsable métier et procédure de reprise	
Cas de référence, versions et procédure de rejeu sans action réelle	
Sources, règles et traces à présenter pour expliquer le résultat	
Seuil de qualité, stabilité attendue et budget à ne pas dépasser	

Avancez par preuves : **observer le travail actuel**, tester hors production, proposer des brouillons, mesurer le gain net, puis élargir seulement le périmètre qui fonctionne. Arrêtez ou revoyez le pilote si les corrections consomment le temps économisé, si les sources sont insuffisantes ou si les données ne peuvent pas être traitées dans de bonnes conditions.

Du développement logiciel à l'automatisation utile

Je suis **Jules Lepoivre, fondateur de Saki Studio**. Depuis près de vingt ans, je développe des logiciels et accompagne des équipes dans leurs projets numériques. L'automatisation par l'IA prolonge ce travail : comprendre un métier, relier les bons outils, fiabiliser l'exécution et transmettre la méthode.

Mon intervention peut aller du cadrage d'un processus à son intégration dans vos logiciels : cadre d'exécution adapté, mémoire utile, tests de répétabilité, traces compréhensibles et formation des personnes qui l'utilisent. Le point de départ est votre travail quotidien, ses règles et ses exceptions. [17, 20]

Orchbun est l'une des briques possibles de ce cadre : un outil ouvert pour lancer des agents et maintenir une mémoire de projet locale. Pour une équipe commerciale ou administrative, Saki Studio adapte les écrans, les connecteurs, le vocabulaire et les validations au travail réel. Les garanties du dispositif dépendent de ce montage complet et des fournisseurs sollicités. [17]

Quel travail aimeriez-vous ne plus refaire chaque semaine ? Répondez à mon message LinkedIn avec la tâche, les outils utilisés et sa fréquence. N'envoyez pas de données confidentielles : un exemple fictif suffit pour amorcer l'échange.

Jules Lepoivre · Saki Studio

Sources

Sources consultées le 9 septembre 2026. Les exemples de volumes et de gains sont illustratifs. Les scores publics ne sont pas des tests de fiabilité métier.

- [01] Anthropic — [Concevoir le contexte utile des agents IA](#)
- [02] Anthropic — [Building effective agents : workflows et autonomie](#)
- [03] CNIL — [Questions-réponses sur l'utilisation d'un système d'IA générative](#)
- [04] CNIL — [Garantir la sécurité du développement d'un système d'IA](#)
- [05] CNIL / CIANum — [IA agentique et données personnelles, note du 20 juillet 2026](#)
- [06] OWASP — [LLM01:2025 — Prompt Injection](#)
- [07] OWASP — [LLM06:2025 — Excessive Agency](#)
- [08] OpenRouter — [Claude Fable 5.1 : tarifs, fournisseurs et avertissement de rétention](#)
- [09] OpenRouter — [GPT-6 Astra : tarifs et fournisseurs](#)
- [10] OpenRouter — [Kimi K3 : tarifs et fournisseurs](#)
- [11] OpenRouter — [MiniMax M3 : tarifs et fournisseurs](#)
- [12] Artificial Analysis — [Classement des modèles — Intelligence Index et configurations](#)
- [13] OpenRouter — [Tarification de la plateforme](#)
- [14] OpenRouter — [Zero Data Retention : portée, paramètres et limites](#)
- [15] OpenRouter — [Données collectées et confidentialité](#)
- [16] n8n — [Validation humaine des appels d'outils IA](#)
- [17] Saki Studio — [Mémoire IA : Orchbun, adaptation aux équipes et limites](#)
- [18] OpenRouter — [Tokens de raisonnement et facturation](#)
- [19] Make — [Documentation des scénarios](#)
- [20] Saki Studio — [Jules Lepoivre : logiciels, équipes et produits](#)
- [21] OpenAI — [Unlocking the Codex harness : orchestration, état, outils et approbations](#)
- [22] Anthropic — [Effective harnesses for long-running agents : contexte, continuité et tests](#)
- [23] Anthropic — [Demystifying evals for AI agents : évaluer le modèle avec son harness et répéter les essais](#)
- [24] Anthropic — [Reasoning models don't always say what they think : limites des justifications affichées](#)
- [25] CNIL — [Les six grands principes du RGPD](#)
- [26] CNIL — [RGPD : les premières étapes](#)
- [27] CNIL — [Ce qu'il faut savoir sur l'analyse d'impact relative à la protection des données](#)
- [28] CNIL — [Profilage et décision entièrement automatisée : article 22 du RGPD](#)
- [29] CNIL — [IA : informer les personnes concernées — fiche consacrée au développement, repères de transparence](#)