IAtus × IAtrogénique

LIVRE BLANC • OCTOBRE 2026

Quand l’IA fait science

Produire plus, savoir mieux ?

Hallucinations, publications et chaîne de preuve en santé



PROPOSITION VALIDATION DÉCISION

Suivre la preuve, de l’hypothèse à la décision.

Une lecture critique des études, des rapports et des usages de l’IA en recherche et en santé.

Lionel Reichardt

PHARMAGEEK

OUVERTURE • CADRAGE

Le contrat de lecture

Ce livre blanc s’adresse aux chercheurs, professionnels de santé, équipes éditoriales, formateurs et responsables d’établissement. Il examine les usages de l’IA depuis la production d’une hypothèse jusqu’à l’utilisation d’une connaissance dans les soins. Son objet est la qualité de la chaîne de preuve : provenance, validité, interprétation, applicabilité et correction.

Les résultats retenus ne décrivent pas une IA unique. Prédiction structurale, assistance documentaire, génération de code et aide à la décision clinique sont des interventions différentes. Leurs performances s’évaluent sur des tâches et des critères distincts. Un modèle peut réussir une tâche bien délimitée et échouer lorsqu’il change de population, de corpus ou de rôle dans le travail.

Trois propositions guident la lecture. Premièrement, produire un objet scientifique ne suffit pas à valider son contenu. Deuxièmement, une vérification doit porter sur les éléments dont dépend une décision, pas seulement sur l’apparence du document. Troisièmement, le gain de temps doit intégrer le travail de contrôle et les conséquences des erreurs.

Le lecteur trouvera des analyses de travaux, sept fiches visuelles détachables et des outils de travail. Les fiches de gouvernance et les procédures organisent la vérification et la décision dans chaque contexte d’usage.

IAtus examine l’écart entre une promesse et son application. **IAtrogénique** examine les effets indésirables possibles d’un outil censé améliorer les soins. Ces deux regards portent une même exigence : rendre l’innovation vérifiable avant de lui confier une décision.

OUVERTURE • CADRAGE

Synthèse pour les décideurs

L’IA ouvre des possibilités réelles de prédiction, d’exploration et d’assistance scientifique. Leur valeur dépend des contrôles extérieurs : confrontation à des données de référence, expérience préclinique, réplication ou étude chez des patients. Ces étapes répondent à des questions différentes. [1, 2, 15]

La croissance documentaire précède les LLM. La série NSF/NCSES recense environ 2,19 millions de publications en sciences et ingénierie en 2014 et 3,53 millions en 2024. Elle compte des articles de revues et des communications de conférences sélectionnés dans Scopus. Le rapport est publié en 2026 ; le dernier point de cette série est 2024. Cette évolution ne mesure ni l’effet causal de l’IA, ni la qualité des travaux. [28]

Les indicateurs récents doivent être lus à leur échelle : 6,9 heures de gain hebdomadaire moyen déclaré dans une enquête de 637 chercheurs aux États-Unis et au Royaume-Uni ; +42 % de soumissions dans une revue de sciences des organisations. Leur portée est celle des protocoles étudiés. [18, 19]

L’intégrité documentaire exige davantage qu’un lien. L’audit du Lancet identifie des références classées fictives dans un sous-ensemble de la littérature biomédicale ouverte, sans établir qu’elles ont été créées par l’IA. L’existence d’une source doit être vérifiée, puis son soutien effectif à l’affirmation et son applicabilité. [7]

En clinique, un taux par phrase ne représente pas un taux de patients lésés. L’expérience d’Asgari porte sur des consultations simulées. La gravité est potentielle et ne remplace pas une mesure de dommages en situation réelle. [9]

Décision proposée : autoriser des usages ciblés, documenter leur périmètre, fournir les moyens de vérification, mesurer le travail complet, puis organiser correction et suspension. Une réputation de modèle ou une démonstration réussie ne suffit pas à transférer la confiance à l’ensemble d’un parcours de soins.

OUVERTURE • CADRAGE

Méthode, corpus et niveau de confiance

Cette synthèse narrative critique croise les travaux de la newsletter d’origine, des études complémentaires et des sources institutionnelles. Le corpus est arrêté au **5 octobre 2026**. Les références privilégient les articles originaux et les versions auteurs identifiées ; les articles de presse servent de contexte et de points d’entrée.

Chaque résultat est examiné à partir de son protocole, de son unité de mesure, de sa population ou de son corpus, et de la conclusion qu’il soutient. Les contre-analyses sont conservées lorsqu’elles modifient l’interprétation. Les dates, dénominateurs et niveaux de validation accompagnent les données.

LE CONTRAT DE PREUVE

Une synthèse critique, pas une revue systématique. La sélection n’est ni exhaustive ni issue d’un protocole enregistré ; elle ne produit pas de méta-analyse.

Des résultats situés. Association, effet causal, gain déclaré et gain mesuré sont distingués. Prédiction, validation préclinique et bénéfice patient répondent à des questions différentes.

Des recommandations assumées. Les grilles, matrices et procédures expriment les recommandations de l’auteur. Elles servent à décider et à organiser les contrôles ; elles ne valent ni norme réglementaire, ni certification, ni validation d’un produit.

Des unités conservées. Aucun taux par phrase, référence ou document n’est converti en pourcentage de patients lésés. Les variations calculées sont identifiées et les exemples de pratique sont hypothétiques.

Un accès documenté. Les limites d’accès figurent dans la bibliographie. L’audit du Lancet est lu dans son texte principal, sans annexe ; le rapport OMS 2026 est mobilisé à partir de sa présentation officielle et du communiqué.

Transparence de préparation : assistance d’IA pour la recherche, la structuration et la mise en page. Les références scientifiques soutiennent les résultats ; la validation finale avant diffusion relève de l’auteur. Les changements de système, de corpus ou de contexte appellent une réévaluation.

SOMMAIRE

Parcours de lecture

Ouverture : contrat, synthèse et méthode 2

Contexte : une science déjà sous pression 6

1–2 • Découverte et agents scientifiques 8

3 • Productivité, volume et qualité 13

4–5 • Sources et interprétation 17

6 • Erreurs et risque clinique 22

7–8 • Recherche documentaire et synthèse 26

9–10 • Publication et évaluation 30

11–12 • Bénéfice patient et supervision 35

13–14 • Gouvernance et six préalables 39

Annexes A–H : outils de travail 45

Sources et repères de lecture 53

Travaillons ensemble • Pharmageek 56

Les sept infographies occupent chacune une page entière. Un livret séparé permet de les consulter et de les imprimer indépendamment. Les signets du PDF donnent accès aux chapitres et aux annexes.

OUVERTURE • CADRAGE

Avant les LLM : une science déjà sous pression

La hausse du nombre de publications n’a pas commencé avec ChatGPT. Ce constat importe pour l’attribution des causes : les incitations de carrière, le financement, la croissance du nombre de chercheurs, les pratiques disciplinaires et les modèles éditoriaux existaient avant la généralisation des assistants génératifs.

La série NSF/NCSES publiée en mai 2026 permet de situer cette croissance sur un périmètre cohérent. Elle porte sur des publications mondiales en sciences et ingénierie indexées dans Scopus : articles de revues évaluées et communications de conférences sélectionnées. Les données ont été extraites en août 2025 et couvrent 2014–2024. Elles ne constituent pas un décompte de tout texte scientifique produit dans le monde. [28]

Entre les deux extrémités de la série, le volume passe de 2 193 498 à 3 525 866, soit **+60,7 %**, calcul à partir des valeurs publiées. Entre 2023 et 2024, la progression est d’environ **+7,3 %**. Ces calculs décrivent une série ; ils ne permettent pas d’isoler l’effet des LLM.

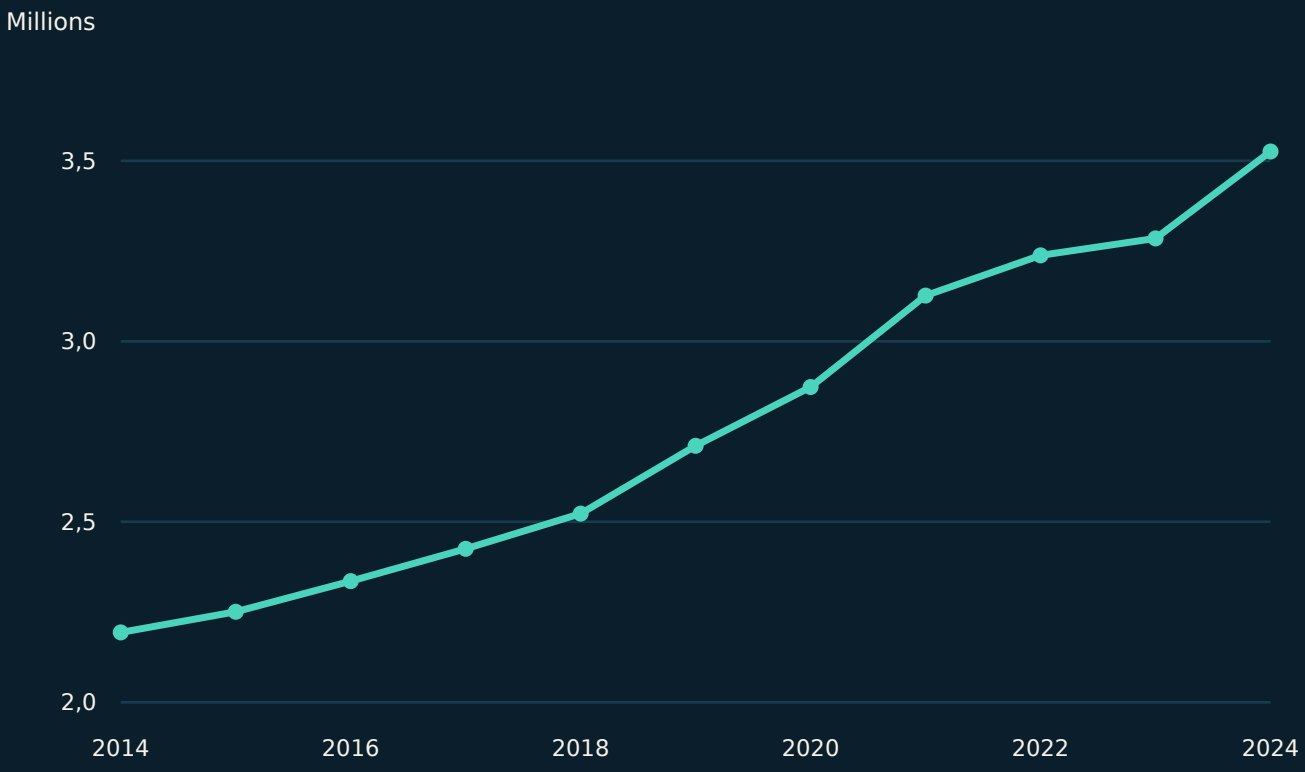
Hanson et ses collègues décrivent également une croissance et des tensions de sélection dans un autre corpus bibliométrique couvrant 2016–2022. Leur travail invite à étudier la capacité de relecture et les transformations du système éditorial, mais ses nombres ne doivent pas être raccordés à la série NSF comme s’il s’agissait d’un même compteur. [27]

Ce que l’on peut conclure : le système reçoit et diffuse un volume documentaire élevé et croissant sur ces périmètres. **Ce que l’on ne peut pas conclure :** chaque publication supplémentaire serait de moindre qualité, ni qu’une progression après 2022 serait nécessairement causée par ChatGPT. Aucune valeur pour 2025 ou 2026 n’est ajoutée à la courbe présentée ici.

La question pour l’IA devient alors organisationnelle : quels travaux accélère-t-elle, et les capacités de sélection, d’expérimentation et de correction suivent-elles ? L’ouverture du débat gagne à partir de ces capacités observables.

Une science déjà sous pression documentaire

Publications mondiales en sciences et ingénierie, 2014–2024.



2014 : 2 193 498 • 2024 : 3 525 866

+60,7 % sur la période ; **+7,3 %** entre 2023 et 2024. Variations calculées à partir du tableau officiel.

Le rapport est de 2026. Le dernier point est 2024.

Périmètre : revues évaluées et communications de conférences sélectionnées, champs sciences et ingénierie, Scopus. Extraction : août 2025. Aucun point 2025 ou 2026 extrapolé.

Source : National Science Board / NCSES, NSF, rapport 2026, figure 29. La courbe ne mesure ni la qualité des articles ni l’effet causal de l’IA.

ANALYSE • ÉTUDES ET PORTÉE

1. Les avancées scientifiques méritent une évaluation précise

Parler de « l’IA dans la science » recouvre des systèmes très différents. Un modèle de prédiction moléculaire, un assistant de recherche bibliographique et un agent qui écrit du code n’ont ni les mêmes objectifs, ni les mêmes données d’entrée, ni les mêmes modes d’échec. Les réunir sous un indicateur général de performance rendrait l’analyse peu utile.

AlphaFold 3 illustre une avancée spécialisée. Publié dans Nature en 2024, le modèle prédit des structures de complexes associant protéines, acides nucléiques, petites molécules et autres composants. Les auteurs rapportent des améliorations sur plusieurs tâches de prédiction d’interactions biomoléculaires. Ils décrivent également un risque de structures plausibles dans des régions désordonnées et les méthodes utilisées pour le réduire. [1]

Une structure prédite aide à choisir une expérience ou à explorer un mécanisme. L’efficacité, la tolérance et l’utilité chez un patient se démontrent dans les étapes expérimentales et cliniques suivantes.

L’étude Co-Scientist, publiée dans Nature en 2026, rapproche génération d’hypothèses et vérification expérimentale. Ce système multiagents propose, critique et affine des pistes de recherche. Les auteurs rapportent notamment des candidats au repositionnement de médicaments et des associations pour la leucémie aiguë myéloïde validés par des expériences **in vitro**. Le niveau de validation obtenu est préclinique. [2]

Ce qui rend ces travaux intéressants est la présence d’un contrôle extérieur à la génération du texte : une mesure, une expérience ou une comparaison à des données de référence. Pour apprécier une promesse d’accélération, je propose de regarder d’abord cette étape. Que produit le système ? Qui vérifie ce résultat ? Et quelle décision cette vérification autorise-t-elle ?

APPROFONDISSEMENT • MISE EN PRATIQUE

Décomposer la promesse de découverte

Une annonce scientifique contient souvent plusieurs promesses emboîtées : trouver un candidat, améliorer une prédiction, réduire le coût d’une expérience et produire un bénéfice utilisable. L’évaluation doit déplier ces promesses avant de choisir un indicateur. Pour une molécule, une meilleure représentation structurale peut aider à organiser des essais ; elle n’établit ni l’exposition nécessaire, ni la tolérance, ni le bénéfice attendu dans une indication.

Exemple hypothétique. Une équipe reçoit vingt candidats générés par un système. Trois montrent une activité dans un modèle cellulaire. Le résultat intéressant porte sur la sélection et sur les mesures expérimentales obtenues, avec leurs contrôles. Il ne peut pas être résumé par « l’IA a trouvé trois traitements ». Les candidats inactifs, les conditions d’essai et les critères de sélection font partie de l’information utile.

Pour apprécier l’accélération, il faut choisir un point de départ et un point d’arrivée comparables : génération initiale, candidat confirmé, reproduction indépendante ou étape réglementaire. Le temps gagné à l’une de ces étapes ne prédit pas automatiquement celui de l’ensemble du développement. Comparer les délais sur des étapes identiques rend le gain interprétable.

Une évaluation robuste devrait aussi comparer la démarche à une stratégie pertinente. Un nombre élevé de candidats ne signifie pas un meilleur taux de validation. Le coût des expériences négatives, la disponibilité des équipements et les compétences nécessaires au suivi peuvent déplacer le bénéfice vers certaines équipes plutôt que le distribuer uniformément.

Enfin, la nouveauté doit être définie. Un rapprochement inédit dans une réponse peut déjà exister dans une publication non retrouvée. Le contrôle documentaire de l’antériorité et la validation expérimentale accomplissent deux fonctions distinctes. AlphaFold 3 et Co-Scientist illustrent des objets scientifiques distincts, à évaluer sur leurs critères propres. [1, 2]

Question	Pièce à demander
Quel objet est produit ?	Structure, candidat, protocole ou résultat : nommer l’unité.
Quel contrôle extérieur ?	Données de référence, expérience, réplication et comparateur.
Quel échec est conservé ?	Résultats négatifs et raisons d’exclusion.
Quelle conclusion reste interdite ?	Énoncer la limite préclinique ou méthodologique.

Les niveaux de validation

Trois études, trois questions : objet, contrôle, conclusion et incertitude.

PRÉDICTION — AlphaFold 3, 2024 [1]

Objet : structures de complexes biomoléculaires.
Contrôle extérieur : comparaison aux données structurales de référence.
Conclusion : performance de prédiction dans les tâches évaluées.
Reste à établir : validation expérimentale et éventuel bénéfice clinique.

PRÉCLINIQUE — Co-Scientist, 2026 [2]

Objet : hypothèses et candidats biomédicaux.
Contrôle extérieur : expériences précliniques sélectionnées avec les chercheurs.
Conclusion : résultats dans ces systèmes expérimentaux.
Reste à établir : reproduction, tolérance et efficacité chez l’humain.

CLINIQUE — Essai au Kenya, 2026 [15]

Objet : aide à la décision en soins primaires.
Contrôle extérieur : essai randomisé en grappes, 16 structures, 9 691 patients recrutés.
Conclusion : pas de différence significative sur le critère principal à 14 jours.
Reste à établir : précision, transférabilité et effets dans d’autres usages.

Sources : Abramson et al., Nature 2024 ; Gottweis et al., Nature 2026 ; Agweyu et al., Nature Medicine 2026. Objets et conclusions distincts.

ANALYSE • ÉTUDES ET PORTÉE

2. Un article généré automatiquement reste un objet à examiner

Les agents scientifiques ajoutent une autre dimension : ils peuvent enchaîner recherche documentaire, formulation d’hypothèses, programmation, exécution d’expériences numériques et rédaction.

Le travail Towards end-to-end automation of AI research, publié dans Nature en 2026, décrit cette approche avec The AI Scientist. Trois manuscrits, sélectionnés par des humains, ont été soumis à l’atelier ICBINB associé à ICLR 2025. Un manuscrit a atteint le seuil d’acceptation ; les soumissions ont été retirées selon le protocole prévu avec les organisateurs. Il faut distinguer ce manuscrit généré de la publication humaine qui décrit et évalue le système, ainsi que l’atelier de la conférence principale. Le domaine étudié est l’apprentissage automatique. [3]

La démonstration porte sur l’apprentissage automatique. L’autonomie en recherche clinique appelle des validations propres à ce domaine.

Pour un laboratoire, l’enjeu devient très concret : conserver les données réellement utilisées, le code exécuté, les versions des outils, les résultats négatifs et les interventions humaines. La reconstitution du parcours exige ces traces au-delà du manuscrit.

Je recommande donc de traiter l’automatisation d’une recherche comme un processus à auditer. La question du nombre de papiers produits doit venir après celle de la qualité des expériences, de la possibilité de les reproduire et de la contribution des résultats aux connaissances existantes.

APPROFONDISSEMENT • MISE EN PRATIQUE

Auditer le parcours d’un agent scientifique

Un agent peut modifier le parcours de recherche tout autant que le manuscrit. Il choisit un jeu de données, exécute un script, inspecte une métrique et décide de poursuivre. Le contrôle doit porter sur ces transitions. Un journal ne devient utile que si une autre personne peut reconstituer les opérations et distinguer les propositions des opérations effectivement réalisées.

Le travail sur The AI Scientist apporte une précision importante : la soumission à l’atelier ICBINB associé à ICLR 2025 résulte d’une sélection humaine. Un manuscrit atteint le seuil d’acceptation ; les soumissions sont ensuite retirées selon le protocole prévu. Cette expérience encadrée ne signifie pas qu’un article autonome a été publié à la conférence principale. Les interventions de sélection et le rôle des organisateurs appartiennent à l’interprétation. [3]

Exemple hypothétique. Un agent supprime des observations qu’il juge aberrantes, relance l’analyse et écrit que l’effet est robuste. Pour vérifier cette affirmation, le laboratoire a besoin des données antérieures à l’exclusion, de la règle appliquée, du script, des sorties et de l’analyse sans cette modification. L’explication rédigée après coup ne remplace pas les traces.

La supervision des actions ajoute un risque de validation routinière. Sous la pression du temps, une personne peut accepter une séquence proposée sans reprendre les vérifications décisives. Nous désignons ici cette situation comme un risque d’inertie de validation : il s’agit d’un repère éditorial, pas d’un biais clinique quantifié dans les études citées.

Je propose de séparer les droits de lecture, d’écriture, d’exécution, de transmission et de destruction. Une autorisation de chercher des articles ne devrait pas autoriser implicitement l’envoi d’un manuscrit, la modification d’une base ou le lancement d’une expérience coûteuse. Pour chaque action sensible, le contrôleur doit voir l’objet concerné, la conséquence et le moyen de retour.

Action	Contrôle proposé
Lire et chercher	Corpus autorisé et trace des requêtes.
Exécuter ou modifier	Code versionné, environnement isolé, différences conservées.
Transmettre	Destinataire, contenu final et autorisation explicite.
Suspendre	Arrêt accessible et état récupérable.

ANALYSE • ÉTUDES ET PORTÉE

3. Davantage de publications ne suffit pas à mesurer le progrès

L’IA peut réduire certains obstacles à la rédaction, à la programmation ou à l’accès à des travaux éloignés de sa spécialité. Pour apprécier l’effet sur la science, il faut cependant choisir avec soin ce que l’on mesure.

Kusumegi et ses collègues, dans Science en décembre 2025, rapportent une augmentation de production chez les chercheurs dont les manuscrits sont détectés comme assistés par LLM, ainsi qu’une diversification des références. Ils examinent également le changement de relation entre sophistication rédactionnelle et publication ultérieure. Cette dernière sert d’indicateur indirect de qualité. [4]

Ces résultats ont reçu une critique méthodologique importante. Renault, Bergeaud et Bosquet montrent que dater l’adoption au premier article détecté relie mécaniquement cette date au volume de production. Leurs simulations et tests placebo font apparaître une hausse sans effet causal de l’IA. Leur contre-analyse porte sur l’identification causale du gain annoncé. [5]

Cette controverse a toute sa place dans une newsletter sur la fiabilité. Une étude peut être réelle, publiée dans une revue prestigieuse et porteuse d’une conclusion qui doit encore être discutée.

Une autre analyse, celle de Hao et ses collègues dans Nature en 2026, associe l’usage d’outils d’IA à davantage de publications et de citations pour les chercheurs, tout en observant un resserrement du champ collectif d’exploration. L’analyse est observationnelle et porte sur l’IA au sens large. [6]

Pour les institutions, ces travaux invitent à évaluer plusieurs dimensions séparément : le volume produit, le temps consacré, la solidité des résultats et les nouvelles questions rendues accessibles. Un laboratoire pourrait publier plus vite tout en restant concentré sur les mêmes problèmes. Il pourrait aussi publier moins et réaliser une avancée majeure. Le nombre d’articles ne permet pas de trancher entre ces situations.

Des heures économisées, mais des goulots d’étranglement déplacés

Le rapport de recherche AI in Science: Early Insights, publié en septembre 2026 par une équipe associant notamment Google, Google DeepMind et MIT FutureTech, apporte un indicateur de temps. L’enquête auprès de **637 chercheurs actifs aux États-Unis et au Royaume-Uni** rapporte près de **sept heures économisées par semaine**. Le gain est déclaré dans une enquête située ; son interprétation repose sur ce mode de mesure et sur l’échantillon recruté. Le rapport relève aussi un stock d’hypothèses non testées et une demande importante de vérification des résultats de l’IA. [18]

Le déplacement du travail est aussi important que son accélération. Une hypothèse peut désormais être proposée en quelques secondes ; les échantillons, le recrutement de participants, les expériences et les arbitrages scientifiques restent soumis à des contraintes matérielles et humaines. La démocratisation de la rédaction mérite donc d’être accompagnée par des moyens de validation.

Une pression éditoriale documentée, à son échelle

Dans Organization Science, Gartenberg et ses collègues rapportent une hausse de **42 % des soumissions** depuis l’arrivée de ChatGPT. Leur analyse couvre les soumissions et les évaluations d’une revue de sciences des organisations. Elle mobilise le détecteur Pangram et des mesures de lisibilité : ces indicateurs ne prouvent ni l’usage d’IA dans chaque texte ni la qualité intrinsèque des résultats scientifiques. Les auteurs observent aussi une évolution des relectures vers des textes moins diversifiés dans leurs thèmes. [19]

Ce résultat documente une tension éditoriale dans la revue étudiée. Son transfert à la biomédecine demande une étude spécifique. Les incitations à publier, les règles de carrière et les capacités de sélection participent à cette tension ; l’outil n’en est pas la seule composante.

Une décision d’arXiv illustre la réponse opérationnelle. Depuis l’annonce du 31 octobre 2025, les articles de synthèse et les textes de position soumis à la catégorie **informatique** doivent avoir été acceptés après une évaluation complète par une revue ou une conférence, avec une référence vérifiable. arXiv explique cette évolution par un afflux devenu difficile à modérer, favorisé par les LLM. La règle cible ces types de textes dans la catégorie informatique. [20]

La qualité doit être mesurée selon plusieurs dimensions

Dimension	Apport possible de l’IA	Ce qu’il faut évaluer
Expression et accès	Aide linguistique, programmation, exploration documentaire	Exactitude conservée et accès effectif aux sources
Validité	Analyse et prédiction spécialisées	Méthodes, données, contrôles et reproductibilité
Originalité	Exploration d’hypothèses et rapprochement de domaines	Contribution nouvelle, diversité des questions et validation extérieure
Utilité clinique	Soutien à une tâche de soin	Bénéfice patient, erreurs, charge de contrôle et contexte d’utilisation

L’amélioration d’une dimension ne garantit pas celle des autres. Une rédaction plus claire peut faciliter l’évaluation ; elle peut aussi rendre une conclusion fragile plus convaincante. Le risque de conformisme doit être étudié au niveau des questions explorées et des résultats obtenus. Le seul fait qu’un modèle ait appris sur la littérature passée ne démontre pas qu’il étouffe nécessairement les découvertes de rupture.

APPROFONDISSEMENT • MISE EN PRATIQUE

Construire un tableau de bord qui mesure le travail

Une institution qui demande « combien d’articles supplémentaires ? » choisit déjà une partie de la réponse. Cette métrique peut être utile pour décrire un flux, mais elle ne représente pas le résultat scientifique final. Il faut lui associer des mesures qui ne dépendent pas uniquement de l’élégance du texte ou de sa probabilité d’acceptation.

La controverse Kusumegi–Renault invite à une discipline élémentaire : vérifier comment l’adoption d’un outil est datée avant d’en interpréter l’effet. Si l’entrée dans le groupe utilisateur dépend de la publication d’un texte détectable, elle peut sélectionner un moment de forte activité. Une courbe avant–après ne corrige pas à elle seule ce problème. [4, 5]

Dans un pilote local, je propose de conserver le temps sans assistance, le temps avec assistance, le temps de contrôle et les reprises. Les tâches et les professionnels doivent être comparables. Une alternance de conditions ou une assignation adaptée peut être envisagée, en tenant compte de l’apprentissage, de la contamination entre groupes et de la sécurité. La méthode doit être choisie avant de consulter les résultats.

Les indicateurs de contenu pourraient inclure la traçabilité des affirmations décisives, le nombre de résultats reproduits, la disponibilité du code et la diversité des questions effectivement testées. La validité de ces indicateurs doit elle-même être discutée : une case remplie ou un dépôt de code ne garantit pas une expérience solide.

L’enquête de 637 chercheurs apporte un signal d’usage, avec un échantillon en ligne non probabiliste aux États-Unis et au Royaume-Uni, interrogé de fin juillet à début août 2026. Le gain moyen déclaré de 6,9 heures ne remplace pas une mesure locale du travail complet. De même, les +42 % de soumissions concernent une seule revue. Les protocoles situent la portée de ces résultats. [18, 19]

Dimension	Indicateur à construire localement
Temps complet	Production + vérification + correction + reprises.
Solidité	Résultats contrôlés ou reproduits, pas seulement textes terminés.
Exploration	Questions effectivement testées et alternatives conservées.
Équité	Accès, coûts, compétences et résultats selon les équipes.

Produire et valider : deux capacités à mesurer

Accélération possible et validation nécessaire : deux mesures distinctes.

RÉDACTION	Vérification des affirmations et des références
HYPOTHÈSES	Expériences et résultats négatifs conservés
ANALYSE	Contrôles, données, code et réplication
SOUMISSIONS	Sélection et relecture scientifique

6,9 h / semaine

Gain moyen **déclaré**, enquête en ligne de **637 chercheurs** USA/Royaume-Uni, juillet–août 2026. Échantillon non probabiliste. Rapport de septembre 2026. [18]

+42 %

Soumissions dans **une seule revue**, Organization Science. Analyse publiée en avril 2026 ; détecteur de texte et indicateurs de lisibilité. Pas une hausse mondiale. [19]

1,47 %

191 / 12 999 phrases avec hallucination : expérience de synthèse de consultations **simulées**, publiée en mai 2025. Pas un taux de patients lésés. [9]

Sources : Codreanu et al., 2026 ; Gartenberg et al., 2026 ; Asgari et al., 2025. Trois mesures distinctes : temps déclaré, soumissions et erreurs par phrase.

ANALYSE • ÉTUDES ET PORTÉE

4. Les références fictives atteignent la littérature biomédicale publiée

L’audit de Topaz et ses collègues, publié dans The Lancet en mai 2026, documente les références fictives à grande échelle. [7]

- **Corpus :** 2 471 758 articles du sous-ensemble ouvert de PubMed Central, de janvier 2023 au 18 février 2026.
- **Références vérifiées :** 97,1 millions portant un identifiant PubMed.
- **Références classées fictives :** 4 046, réparties dans 2 810 articles.
- **Validation :** précision estimée à 91 % ; rappel non mesuré.
- **Périmètre :** références sans PMID exclues ; sous-ensemble ouvert de la littérature biomédicale.

L’audit identifie les fabrications ; leur origine appelle une investigation distincte. Les auteurs envisagent l’usage non critique de l’IA, les activités de paper mills et la fraude. Ces causes demandent chacune des éléments de provenance.

Une référence problématique exige ensuite un examen de son rôle. Soutient-elle une phrase de contexte ou une conclusion essentielle ? La correction suffit-elle ? La validité des résultats dépend-elle d’elle ? Ces questions guident la réponse éditoriale et l’évaluation scientifique.

Priorité opérationnelle : vérifier l’identité des références, contrôler les affirmations qu’elles soutiennent et retrouver les documents qui ont repris la conclusion.

APPROFONDISSEMENT • MISE EN PRATIQUE

Réagir à une référence classée fictive

La découverte d’une référence problématique ne dispense pas d’examiner la chaîne qui l’a introduite. Elle peut provenir d’une génération automatique, d’une erreur de gestion bibliographique ou d’une fabrication intentionnelle. Le diagnostic doit être étayé avant d’attribuer une cause ou une intention.

L’audit du Lancet utilise une procédure automatisée, comprenant un modèle de langage et des recherches dans plusieurs bases, puis une validation humaine d’un échantillon. Cela montre que l’IA peut aussi participer au contrôle. Cela ne rend pas le classement infaillible : la précision mesurée doit être conservée, et l’absence d’évaluation du rappel laisse inconnues les fabrications non détectées. [7]

Exemple hypothétique. Un PMID mène à un article de dermatologie, alors que le manuscrit lui attribue un essai cardiologique. Le contrôleur vérifie d’abord s’il s’agit d’une erreur d’identifiant ou d’un article introuvable. Il recherche ensuite le passage censé soutenir l’affirmation. Enfin, il apprécie si la conclusion principale dépend de cette citation. Ces étapes peuvent conduire à des réponses différentes.

Je propose une gestion graduée : conserver la référence originale et le constat, demander une source identifiable aux auteurs, puis documenter l’effet sur les conclusions. Une correction bibliographique peut suffire lorsque le travail existe et que son contenu soutient correctement l’affirmation. Une source inexistante portant une conclusion essentielle appelle une réévaluation plus profonde.

Une correction ne doit pas seulement modifier le fichier courant. Les versions déjà diffusées, les synthèses internes et les supports de formation peuvent avoir repris l’erreur. L’organisation doit définir où placer l’avis de correction et comment retrouver les utilisateurs concernés. Le volume des reprises est un objet d’enquête ; aucun taux universel de propagation ne peut être déduit de cet exemple.

Étape	Trace à conserver
Identifier	Titre, auteurs, identifiants et version reçue.
Vérifier	Bases consultées, résultats et date du contrôle.
Qualifier	Erreur de métadonnées, source non retrouvée ou fabrication étayée.
Corriger	Effet sur l’affirmation et liste des documents concernés.

ANALYSE • ÉTUDES ET PORTÉE

5. Une source peut exister et être mal interprétée

Vérifier un titre, un DOI ou un PMID est nécessaire. Le contrôle doit se poursuivre dans le contenu.

Prenons un **exemple hypothétique** : un assistant cite correctement une étude menée sur des cellules et écrit que le traitement a démontré son efficacité chez l’humain. La source est réelle. L’affirmation dépasse pourtant ce qu’elle établit.

D’autres glissements demandent la même attention : transformer une association en causalité, appliquer un résultat à une population exclue de l’étude, confondre un critère biologique avec une amélioration clinique ou présenter une absence de différence significative comme une équivalence démontrée.

Le commentaire When machines misread science, publié dans npj Digital Medicine en 2026, propose notamment des dispositifs pour expliciter les limites d’interprétation des publications par les machines et les humains. L’évaluation de ces propositions est une étape de recherche à poursuivre. [8]

La chaîne de contrôle que je recommande comporte donc trois questions : **la source existe-t-elle ? Soutient-elle l’affirmation ? Cette affirmation est-elle applicable à la décision envisagée ?**

Cette dernière question exige un jugement contextuel. Une preuve peut être solide dans son cadre et insuffisante pour décider dans un autre. L’ajout de citations ne résout pas automatiquement cette difficulté.

APPROFONDISSEMENT • MISE EN PRATIQUE

L’affirmation comme unité d’audit

Un document peut contenir des passages fiables et une seule inférence décisive insuffisamment soutenue. L’unité pratique de contrôle est donc souvent l’affirmation : ce qui est avancé, à quel endroit et pour quelle décision. L’audit doit distinguer la citation de l’inférence ajoutée par l’auteur ou par le modèle.

Je propose un registre comprenant la formulation exacte, la source, le passage d’appui, la population, le critère, l’incertitude et la limite de transfert. Une phrase telle que « améliore la prise en charge » est trop large si l’étude mesure seulement la lisibilité d’un compte rendu. Le registre impose de retrouver le résultat réellement observé.

Exemple hypothétique. Une étude rapporte une association entre l’usage d’un assistant et une meilleure complétude des notes. Une synthèse conclut que cet assistant réduit les événements indésirables. Le premier résultat pourrait être correct ; le second exige une autre mesure. Le contrôle ne consiste pas à trouver davantage de liens, mais à identifier cette transition.

Le choix des mots est une partie du contrôle. « Compatible avec », « associé à », « prédit », « observé dans ce protocole » et « démontré dans cet essai » portent des engagements différents. Une reformulation linguistique peut modifier ces engagements même lorsqu’elle conserve les nombres.

Pour un usage clinique, une affirmation générale doit ensuite rencontrer les caractéristiques du patient et de l’organisation : âge, comorbidités, exclusions, ressource disponible, alternative et préférence. La traçabilité documentaire aide au jugement ; elle ne l’automatise pas. Les propositions de guardrails de Peoples et ses collègues appartiennent à ce débat, sans constituer un standard validé. [8]

Le registre n’a pas vocation à alourdir chaque phrase d’un article. Il vise les assertions qui portent une conclusion, une recommandation ou une décision. Son périmètre doit être proportionné à l’usage et au risque.

Glissement	Question corrective
Association → cause	Le protocole identifie-t-il un effet causal ?
Critère intermédiaire → bénéfice patient	Un résultat clinique a-t-il été mesuré ?
Résultat moyen → patient particulier	La population et les exclusions permettent-elles ce transfert ?
Non-significatif → équivalent	Un protocole d’équivalence ou de non-infériorité était-il prévu ?

Les trois contrôles d’une affirmation

Trois questions successives. Exemples explicitement hypothétiques.

01 — La source existe-t-elle ?

Un DOI fonctionne, mais renvoie à un autre article. Vérifier ensemble titre, auteurs, année et identifiant.

02 — Soutient-elle l’affirmation ?

Une étude sur des cellules est citée comme preuve d’efficacité chez l’humain. Lire le passage et le protocole : la source réelle ne soutient pas cette conclusion.

03 — Le résultat s’applique-t-il à la décision ?

Une étude exclut une population à laquelle appartient le patient. Une interprétation correcte ne suffit pas à autoriser ce transfert.

Existence ≠ soutien ≠ applicabilité

Source : Peoples et al., npj Digital Medicine, 2026 ; lecture critique développée dans ce livre. Aucun incident réel illustré.

ANALYSE • ÉTUDES ET PORTÉE

6. En clinique, le dénominateur et la gravité comptent

Repères — six phénomènes à distinguer

Terme	Sens retenu dans cette édition	Exemple ou conséquence
Hallucination intrinsèque	Le résultat contredit le document source.	Une note inverse une information sur un traitement.
Hallucination extrinsèque	Le résultat ajoute une information non étayée par la source. Elle n’est pas nécessairement fausse dans l’absolu.	Une note attribue au patient un antécédent absent du dossier.
Omission	Une information nécessaire est absente du résultat.	Une allergie documentée disparaît de la synthèse.
Sycophanie	Le modèle privilégie l’accord avec l’utilisateur au détriment de l’exactitude.	Une hypothèse suggérée reçoit une validation insuffisamment critique.
Biais	Une différence systématique affecte la représentation ou la performance.	Un système est moins fiable pour une population donnée.
Fraude	Une conduite intentionnelle trompe sur les données, les méthodes ou les résultats.	Des données inventées sont présentées comme observées.

Les exemples sont hypothétiques. La distinction intrinsèque–extrinsèque porte ici sur la fidélité aux documents fournis, avec ou sans RAG. La fabrication d’une référence inexistante est un cas d’invention ; toute addition non étayée n’est pas une référence fictive. Le terme « confabulation » est parfois proposé pour éviter l’anthropomorphisme d’« hallucination », mais il ne constitue pas une catégorie clinique à appliquer aux machines.

Sharma et ses collègues documentent la sycophanie dans plusieurs tâches de génération avec des assistants ajustés à partir de préférences humaines. Ces expériences caractérisent le mécanisme de sycophanie dans les tâches étudiées. [26]

Ces distinctions appellent des contrôles différents : fidélité, exhaustivité, confrontation des hypothèses, évaluation par population ou investigation d’intégrité.

Un « taux d’hallucination » n’a de sens qu’avec une tâche, une définition de l’erreur et un dénominateur. Il peut désigner une proportion de références, de phrases, de réponses ou de documents. Ces mesures ne se comparent pas directement.

Asgari et ses collègues ont étudié la synthèse de consultations dans npj Digital Medicine en mai 2025. [9]

- **Entrées :** 25 transcriptions simulées.
- **Configurations :** 18, appliquées aux mêmes transcriptions ; 450 notes produites.
- **Phrases générées :** 12 999, dont 191 avec une hallucination, soit 1,47 %.
- **Gravité potentielle :** 84 des 191 hallucinations, environ 44 %, classées majeures, susceptibles de modifier le diagnostic ou la prise en charge si elles restent non corrigées.
- **Cadre de mesure :** expérience simulée ; aucun dommage patient mesuré ; affiliation Tortus AI à prendre en compte.

Le taux par phrase et la proportion d’erreurs majeures répondent à deux questions différentes. Une fréquence limitée peut rester préoccupante si l’erreur concerne une information décisive.

Il faut aussi distinguer l’ajout d’une information inventée de l’omission d’une information présente. Une note peut contenir uniquement des phrases exactes et rester dangereusement incomplète. La priorité d’évaluation dépend donc de l’usage : pour un compte rendu, l’exhaustivité des informations nécessaires à la continuité des soins doit être examinée avec la fidélité du texte.

Pour un établissement, je recommande un protocole qui mesure à la fois la fréquence des erreurs, leur gravité, leur détection par les professionnels et le temps nécessaire pour les corriger. Le pourcentage seul ne décrit pas la sécurité du processus.

APPROFONDISSEMENT • MISE EN PRATIQUE

Passer du nombre d’erreurs à la sécurité du processus

La fréquence et la gravité ne répondent pas à la même question. Un évaluateur doit encore établir si l’erreur arrive jusqu’à l’utilisateur, si elle est détectée, corrigée et susceptible de modifier une action. Ces maillons ne doivent pas être remplacés par un pourcentage unique.

Dans l’expérience d’Asgari, les 25 transcriptions sont simulées et réutilisées dans plusieurs configurations. Les 450 notes ne représentent donc pas 450 consultations indépendantes. Les 12 999 phrases générées ne sont pas non plus 12 999 patients. Ces dépendances limitent l’interprétation d’une fréquence et toute extrapolation à une activité hospitalière. [9]

Exemple hypothétique. Une synthèse ajoute un traitement jamais prescrit. Si le professionnel l’identifie avant signature, l’incident reste une erreur interceptée, avec un coût de correction. Si le traitement est repris dans une transmission puis administré, la trajectoire devient différente. Une évaluation doit distinguer la sortie erronée, la décision affectée et l’événement patient.

Je propose de mesurer les ajouts non étayés et les omissions sur des unités définies à l’avance, puis de rapporter les résultats au niveau du document et du processus. Une note sans ajout peut rester incomplète ; une note avec un ajout mineur ne présente pas nécessairement un risque majeur. Les critères d’annotation et les désaccords entre évaluateurs doivent être conservés.

La gravité potentielle devrait être évaluée avec des professionnels compétents pour la tâche, autant que possible sans connaître la condition de génération. Le protocole local doit tester la capacité de cette organisation à sécuriser le processus. Les cas rares et décisifs peuvent nécessiter un échantillonnage ciblé en complément d’un corpus représentatif.

Enfin, il faut éviter d’appliquer une formule d’indépendance pour transformer un taux par phrase en risque par patient. Les erreurs peuvent être corrélées, les notes différer en longueur et les contrôles humains intervenir avant toute action.

Unité	Mesure distincte
Phrase ou élément	Fidélité et information non étayée.
Document	Présence d’erreurs et exhaustivité utile.
Processus	Détection, correction, reprises et temps.
Patient	Décision affectée et résultat clinique mesuré.

Lire correctement les chiffres d’erreur

Deux protocoles, deux objets : conserver chaque dénominateur.

AUDIT LANCET — Mai 2026 [7]

Corpus : 2 471 758 articles ouverts PMC, janvier 2023–18 février 2026.
97,1 millions de références avec PMID vérifiées.
4 046 références classées fictives, dans **2 810** articles.
Précision estimée : **91 %** ; rappel non mesuré. Références sans PMID exclues. Ni attribution causale à l’IA, ni mesure de préjudice patient.

ASGARI — Mai 2025 [9]

25 transcriptions simulées × 18 configurations = 450 notes.
191 / 12 999 phrases = 1,47 % avec hallucination.
84 / 191 hallucinations ≈ 44 % classées majeures : gravité potentielle si non corrigées.
Configurations répétées ; aucun dommage patient mesuré. Affiliation Tortus AI à prendre en compte.

Deux protocoles : qualité bibliographique et fidélité de synthèse simulée.

Sources : Topaz et al., Lancet, 2026 ; Asgari et al., npj Digital Medicine, 2025. Texte principal de l’audit consulté ; annexe non consultée.

ANALYSE • ÉTUDES ET PORTÉE

7. La recherche documentaire améliore les systèmes sans terminer la vérification

Le retrieval-augmented generation, ou RAG, associe une recherche dans des documents à la génération d’une réponse. Cette architecture donne au système accès à des sources qu’il peut citer et à partir desquelles il peut travailler.

OpenScholar, présenté dans Nature en février 2026, utilise un corpus de 45 millions d’articles ouverts et un protocole d’évaluation multidisciplinaire. Les résultats montrent l’intérêt d’un système spécialisé de recherche et de synthèse par rapport à des modèles généralistes dans les tâches étudiées. L’évaluation clinique doit porter sur la tâche médicale et la décision envisagée. [10]

Pour l’utilisateur, la bonne interface devrait permettre de retrouver le passage qui soutient chaque affirmation importante. Il faut également savoir quelles publications sont accessibles au système et lesquelles restent hors de son corpus.

Je recommande d’évaluer la recherche documentaire et la génération séparément. Un assistant peut manquer une étude pertinente, puis produire une synthèse fidèle des documents qu’il a trouvés. Il peut aussi retrouver la bonne étude et en tirer une conclusion excessive. Ces deux échecs appellent des corrections différentes.

Un second modèle peut signaler des problèmes ; le contrôle décisif revient aux données ou à la source. Le rôle de l’outil de vérification doit être défini, mesuré et limité à ce qu’il sait effectivement vérifier.

APPROFONDISSEMENT • MISE EN PRATIQUE

Séparer recherche, extraction et génération

Une réponse ancrée dans des documents peut échouer à trois niveaux. La recherche peut manquer une source pertinente ; l’extraction peut déformer un passage trouvé ; la génération peut ajouter une interprétation excessive. Un protocole qui n’évalue que la réponse finale ne permet pas toujours de savoir quelle partie corriger.

Je recommande de conserver un ensemble de questions de référence, les documents attendus et les passages d’appui. Le corpus doit préciser les langues, dates, types de publications et conditions d’accès. Les textes retirés, corrigés ou remplacés doivent pouvoir être signalés. Un corpus uniquement ouvert peut être utile, mais il ne représente pas automatiquement toute la littérature pertinente.

Exemple hypothétique. L’assistant retrouve trois petits essais positifs et manque une grande étude négative. Sa synthèse des trois documents est fidèle. La défaillance porte néanmoins sur la couverture documentaire. À l’inverse, il peut retrouver tous les essais puis attribuer à chacun une conclusion que les résultats ne soutiennent pas.

L’interface devrait montrer les passages, leur provenance et la date, plutôt qu’un simple lien global en fin de réponse. Une citation associée à toute une page laisse au lecteur le travail de trouver ce qui est réellement soutenu. Le contrôle doit porter sur les affirmations décisives et sur les exclusions documentaires.

OpenScholar démontre l’intérêt d’un système spécialisé dans les tâches évaluées. Son usage dans une décision clinique exige une évaluation supplémentaire. Les études historiques de fabrication bibliographique, comme Walters et Wilder, concernent également des tâches et versions précises ; elles ne décrivent pas la performance actuelle de tout système RAG. [10, 31]

Un vérificateur automatique peut être utilisé comme signal d’alerte. Pour que son contrôle soit extérieur au générateur, il doit s’appuyer sur des éléments indépendants et vérifiables. Faire approuver une réponse par un second modèle ne suffit pas à établir son exactitude.

Composante	Évaluation proposée
Recherche	Sources pertinentes trouvées et manquées.
Extraction	Fidélité du passage et du nombre récupérés.
Synthèse	Soutien de chaque affirmation et alternatives.
Actualisation	Corrections, retraits et nouvelles versions visibles.

ANALYSE • ÉTUDES ET PORTÉE

8. La synthèse des preuves engage des choix méthodologiques

Une revue systématique demande de définir une question, chercher les études pertinentes, appliquer des critères d’inclusion, extraire les données et apprécier leurs limites. La rédaction de la synthèse intervient dans ce processus ; elle ne le remplace pas.

Rupa Sarkar, rédactrice en chef de Cochrane, rappelle dans une tribune de Nature publiée en mai 2026 que l’organisation teste l’IA pour améliorer l’efficacité de ses revues, tout en jugeant les outils actuels insuffisants pour remplacer toutes les tâches méthodologiques. Cette tribune apporte un témoignage expert sur les exigences de la synthèse méthodologique. [11]

Pour une équipe de recherche, la question utile consiste à identifier les tâches où l’aide est démontrée et celles où un jugement expert reste nécessaire. Le tri initial de documents, l’extraction d’un tableau et l’appréciation du risque de biais ne doivent pas recevoir une autorisation d’usage indistincte.

Dans une synthèse destinée à éclairer les soins, je recommande de conserver les critères de sélection, les exclusions, les données extraites et les arbitrages humains. Le lecteur doit pouvoir comprendre comment la conclusion a été obtenue et la contester sur des éléments précis.

APPROFONDISSEMENT • MISE EN PRATIQUE

Maintenir une méthode de synthèse contestable

Une revue systématique vise une question définie et applique une méthode explicite. Une synthèse générative peut être utile pour explorer un sujet ; elle ne devient pas une revue systématique parce qu’elle adopte ses rubriques. Les critères de sélection, la recherche documentaire et les décisions d’inclusion doivent exister avant la rédaction finale.

PRISMA 2020 propose un cadre de compte rendu pour les revues systématiques. Son respect améliore la transparence du rapport ; il ne certifie pas l’absence de biais ni la validité de chaque décision méthodologique. Nous utilisons ce repère pour distinguer le processus de la seule apparence formelle. [29]

L’aide par IA devrait recevoir une autorisation par tâche. Pour un tri documentaire, l’équipe doit examiner les études exclues à tort et le dispositif de rattrapage. Pour une extraction, elle doit vérifier les groupes, unités, dates et critères. Pour une appréciation du risque de biais, elle doit conserver les justifications et les arbitrages experts. Aucun bon score sur l’une de ces tâches n’autorise automatiquement les autres.

Exemple hypothétique. Un assistant extrait le nombre d’événements dans un groupe, mais prend l’effectif recruté pour l’effectif analysé. Le calcul peut sembler correct alors que son dénominateur est faux. La vérification revient au tableau original et au parcours des participants, pas à la cohérence interne du texte généré.

La responsabilité de la synthèse implique aussi le traitement des versions multiples d’une même étude, des populations chevauchantes et des résultats non publiés. Les outils doivent permettre de conserver ces liens plutôt que multiplier artificiellement les preuves.

Je propose une double trace : la donnée extraite et son emplacement dans la source, puis la décision prise à partir de cette donnée. Les désaccords doivent être résolus de manière explicite.

Tâche	Trace nécessaire
Recherche	Requêtes, bases, dates et restrictions.
Sélection	Critères et motifs d’exclusion.
Extraction	Valeur, unité, groupe, source et vérification.
Synthèse	Méthode, incertitude et arbitrages documentés.

ANALYSE • ÉTUDES ET PORTÉE

9. Les revues et les institutions doivent préserver une responsabilité identifiable

L’utilisation de l’IA en publication touche la responsabilité des auteurs et la confidentialité de l’évaluation.

Les recommandations de l’ICMJE prévoient la déclaration des usages pertinents et maintiennent la responsabilité humaine du contenu. Une IA ne satisfait pas les critères de qualité d’auteur. Les règles applicables aux relecteurs comportent aussi des exigences de confidentialité : il faut vérifier la politique de la revue avant de transmettre un manuscrit à un service tiers. [12]

Les politiques des éditeurs ne sont pas toutes identiques. Elles distinguent notamment l’aide à la rédaction, les tâches de soutien et le jugement scientifique ; certaines prévoient des exceptions pour la correction linguistique. L’autorisation d’usage doit être vérifiée selon le rôle et les instructions de la revue. [13]

Pour les équipes éditoriales, je propose de concentrer les contrôles sur des objets vérifiables : existence des références, cohérence des données, provenance des figures, disponibilité des méthodes et réponse des auteurs aux anomalies. Un texte très fluide ne prouve pas sa validité ; un indice de style ne suffit pas à établir une fraude.

La déclaration d’un usage d’IA apporte de la transparence. Elle doit être suivie d’un examen de la contribution réelle de l’outil et de la manière dont cette contribution a été vérifiée.

Rétractations : ne pas confondre le sujet d’un article et son mode de fabrication

La revue bibliométrique de Sridharan et Sivaramakrishnan, publiée en janvier 2026, analyse **335 articles rétractés portant sur l’IA**. La sélection porte sur le sujet des articles : l’IA. Les éditeurs initient 98,5 % des rétractations recensées. Le délai médian de 47,5 jours concerne la soumission à l’acceptation ; le délai médian entre publication et rétractation est de 550 jours. Une appréciation de la qualité de la relecture exige une comparaison à des articles non rétractés. [21]

La provenance des défaillances demande un examen spécifique. Par ailleurs, les « tortured phrases » étudiées par Cabanac et ses collègues dès 2021 signalent des reformulations suspectes, antérieures à ChatGPT. Elles servent de signal pour engager une investigation. [25]

APPROFONDISSEMENT • MISE EN PRATIQUE

Transparence, intégrité et responsabilité humaine

Une déclaration d’utilisation de l’IA renseigne la provenance ; l’évaluation de fiabilité et l’investigation d’intégrité examinent ensuite les pièces. L’enjeu éditorial est de comprendre quelles opérations ont été confiées à l’outil et comment leurs résultats ont été vérifiés. La correction linguistique, la création d’une figure et l’analyse de données n’engagent pas le même contenu.

Les politiques ICMJE et Elsevier maintiennent une responsabilité humaine identifiable et fixent des règles de déclaration et de confidentialité. Les dispositions applicables dépendent du rôle, de l’usage et des instructions de la revue ; le traitement d’un manquement repose sur son effet et sur la politique éditoriale. [12, 13]

Je propose une déclaration précise : nom de l’outil, version si disponible, tâches réalisées, parties concernées et modalités de vérification. Une phrase générale telle que « IA utilisée pour améliorer le manuscrit » ne permet pas de savoir si l’outil a modifié une analyse ou simplement corrigé l’orthographe.

Exemple hypothétique. Une figure illustre un mécanisme et ne prétend pas représenter une observation. Sa génération peut être déclarée comme illustration. Si la même image est présentée comme une micrographie issue d’une expérience, la provenance est falsifiée. Le problème n’est pas la seule présence de données synthétiques, mais l’engagement de vérité attaché à leur présentation.

Les détecteurs de style ou de texte généré ne devraient pas être traités comme des preuves individuelles suffisantes. Une investigation doit examiner les données, les méthodes, les figures et les réponses des auteurs. Les reformulations suspectes décrites dès 2021 appellent un examen de provenance. [25]

Le choix entre correction, expression de préoccupation et rétractation dépend de la politique applicable et de l’effet du problème sur la fiabilité du travail. L’investigation documente ces objets avant la décision éditoriale.

Objet	Question d’intégrité
Données	Observées, simulées ou générées : provenance explicite ?
Figures	Illustration ou résultat expérimental ?
Analyse	Code et transformations disponibles ?
Déclaration	Contribution de l’outil et contrôle humain identifiables ?

INTÉGRITÉ ÉDITORIALE • DOSSIER DE PROVENANCE

Marquer les contenus synthétiques

Les comités éditoriaux devraient faire du marquage et de la traçabilité des données synthétiques et des figures générées une exigence de soumission. La déclaration visible doit préciser leur statut, l’outil utilisé, les transformations et les modalités de contrôle. Un tatouage numérique (watermarking) doit être demandé lorsqu’une technique adaptée est disponible, avec un moyen de vérification.

Le watermarking inscrit un signal d’origine dans un contenu. Les métadonnées enregistrent son historique ; les mécanismes de signature peuvent protéger l’intégrité de ce dossier de provenance. Ces techniques ont des fonctions complémentaires. Le rapport NIST AI 100-4 décrit leur intérêt et leurs limites : un marquage peut être supprimé ou falsifié, et sa robustesse varie avec les transformations. Un signal d’origine renseigne la fabrication du contenu, pas sa validité scientifique. [33]

Pour une figure, le comité demande le fichier original, la légende de statut, l’historique des transformations et la méthode de contrôle du marquage. Pour des données tabulaires synthétiques, il demande un dictionnaire des variables, le procédé de génération, les versions et une fiche de provenance associée au jeu de données. Le marquage doit respecter l’usage scientifique ; toute modification des valeurs doit être documentée et validée.

Règle d’examen : la présence d’un watermark est un indice de provenance à contrôler ; son absence appelle l’examen du dossier source. La détection s’appuie sur plusieurs pièces convergentes.

Pièce à demander	Vérification éditoriale
Déclaration et légende	Statut : observation, simulation ou génération ; fonction de la figure
Fichiers et versions	Original, code ou procédé, transformations et responsable
Marquage disponible	Méthode, outil de vérification et robustesse sur les formats diffusés
Dossier de provenance	Métadonnées, liens aux fichiers, contrôle des modifications

Cette exigence place le contrôle à la soumission et le prolonge jusqu’à la diffusion. Elle complète l’investigation des anomalies et l’examen des données.

ANALYSE • ÉTUDES ET PORTÉE

10. Quand l’IA évalue la recherche, elle peut aussi modifier les incitations

L’évaluation par IA peut modifier la sélection des travaux, les décisions de financement et les comportements des chercheurs. Comprendre ce que mesure un score est donc un préalable à son usage.

Dans son analyse disponible sur arXiv en juin 2025, Mike Thelwall compare les indicateurs produits par des LLM aux indicateurs bibliométriques et discute leurs corrélations avec des jugements humains. Il formule un risque d’incitation : une évaluation centrée sur les résumés pourrait pousser les auteurs à exagérer leurs contributions pour améliorer leur score. Cette hypothèse appelle une évaluation des comportements et des décisions. [14]

Pour les établissements et financeurs, le contrôle doit porter sur les critères, les pièces du dossier et la possibilité de contestation. Une aide à la lecture et un score qui détermine effectivement une carrière engagent des responsabilités différentes. Plus la conséquence est importante, plus les éléments vérifiables doivent être accessibles.

Contamination documentaire et model collapse

MODEL COLLAPSE — EFFONDREMENT DU MODÈLE
Dégénérescence statistique qui peut survenir lorsqu’un modèle est réentraîné, de génération en génération, sur des données synthétiques issues des modèles précédents, avec une perte d’information sur les cas rares : les « queues de distribution ». [22]

Shumailov et ses collègues montrent dans Nature en 2024 une perte d’information lors d’entraînements récursifs sur des données générées, notamment dans les parties rares d’une distribution. Le mécanisme dépend des conditions d’entraînement. [22]

La contamination documentaire suit une autre trajectoire : un contenu erroné est publié, repris dans des synthèses, puis éventuellement réutilisé dans un corpus d’entraînement. Sa démonstration exige de retrouver les reprises, leur rôle dans les conclusions et les données effectivement utilisées par le système.

Deux contrôles distincts : retracer la circulation de l’erreur dans les documents ; examiner la composition et le renouvellement des données entre générations de modèles.

Les données synthétiques ont des usages scientifiques légitimes lorsque leur statut est explicite et adapté à l’objectif. Présenter une génération comme une observation falsifie la provenance. Conserver la distinction entre observation, simulation et génération permet d’enquêter sur chaque mécanisme.

APPROFONDISSEMENT • MISE EN PRATIQUE

Protéger la diversité sans l’inventer

La standardisation du style n’équivaut pas à la standardisation de la recherche. Pour examiner un éventuel conformisme, il faut regarder les questions choisies, les alternatives testées, les références mobilisées et les résultats abandonnés. La seule connaissance du corpus d’entraînement ne prouve pas une baisse d’originalité.

Messeri et Crockett analysent les illusions de compréhension et les risques de monocultures associés aux usages d’IA en science. Ce travail conceptuel définit un cadre de vigilance sur les mécanismes de connaissance. Il invite à interroger les façons de produire et de reconnaître la connaissance. [32]

Je propose aux équipes de conserver les hypothèses concurrentes, les arguments qui les départagent et les expériences susceptibles de réfuter la piste privilégiée. Une réponse critique peut être demandée à un assistant, mais le contenu de la contradiction doit ensuite être contrôlé. Plusieurs agents partageant les mêmes sources ne constituent pas nécessairement plusieurs regards indépendants.

La sycophanie renforce ce besoin : un outil peut se rapprocher de l’opinion exprimée par l’utilisateur. Un protocole devrait donc limiter l’injection prématurée de la conclusion souhaitée et comparer les réponses obtenues à des formulations neutres. L’évaluation clinique de cette pratique reste à conduire. [26]

Les scores d’évaluation créent aussi des incitations. Si un résumé convaincant est récompensé, les auteurs peuvent optimiser la formulation plutôt que la contribution. La contestation d’une décision exige alors l’accès aux critères et aux éléments du dossier, pas uniquement un score. [14]

Enfin, contamination documentaire et model collapse doivent rester séparés. Le premier problème exige de retrouver les reprises d’une erreur ; le second concerne les conditions d’entraînement sur des données générées. Les réunir sous une métaphore d’effondrement empêche de choisir le contrôle adapté. [22]

Risque à examiner	Contrôle proposé
Accord automatique	Comparer formulations neutres et orientées.
Pistes trop proches	Conserver alternatives et critères de réfutation.
Score opaque	Rendre visibles critères et possibilité de recours.
Corpus contaminé	Gérer provenance, versions et corrections.

ANALYSE • ÉTUDES ET PORTÉE

11. Le bénéfice pour les patients se mesure dans des études cliniques

Les performances sur des questions médicales ou des cas simulés peuvent orienter la recherche. Le passage à la pratique demande une autre évaluation : celle du système utilisé par des professionnels, avec des patients et dans une organisation réelle.

Un essai pragmatique randomisé en grappes, publié dans Nature Medicine en juin 2026, étudie une assistance par LLM dans 16 structures de soins primaires au Kenya. Il inclut **9 691 patients**. Le critère principal, un ensemble d’échecs de traitement dans les 14 jours, ne diffère pas significativement entre les groupes. Aucun événement indésirable grave n’est jugé lié à l’intervention. L’interprétation porte sur le système étudié et son contexte de soins. [15]

Cette étude rappelle pourquoi les résultats techniques, l’amélioration des processus et les effets sur les patients doivent être mesurés séparément.

Pour la décision d’adopter un outil, je propose de définir le bénéfice recherché avant le déploiement. S’agit-il de temps de documentation, d’exhaustivité d’un compte rendu, d’accès à une expertise ou d’amélioration d’un résultat clinique ? Chaque objectif appelle un protocole et des critères adaptés.

APPROFONDISSEMENT • MISE EN PRATIQUE

Démontrer l’utilité dans un contexte de soins

Un essai clinique examine un système, des professionnels, une organisation et une population. Le produit étudié n’est pas seulement le modèle : il comprend l’interface, les sources accessibles, la formation, les droits d’action et les règles d’escalade. Modifier une de ces composantes peut changer le résultat.

L’essai au Kenya porte sur une assistance dans des soins primaires et un critère patient à 14 jours. Le recrutement total et les effectifs analysés pour le critère principal ne sont pas identiques. L’absence de différence statistiquement significative ne démontre pas une équivalence : la précision de l’estimation et le protocole prévu doivent être examinés. [15]

Je propose de définir avant l’étude le bénéfice recherché et le niveau de preuve nécessaire. Un outil de préparation documentaire peut être évalué sur la fidélité, l’exhaustivité et le temps total. Un outil susceptible de modifier le choix thérapeutique demande une évaluation des décisions et de leurs effets adaptée à ce rôle.

DECIDE-AI propose un cadre de compte rendu des évaluations cliniques précoces d’aide à la décision par IA. Il contribue à rendre visible le contexte humain et organisationnel. Le cadre s’articule avec le protocole d’évaluation et les exigences applicables. [30]

Exemple hypothétique. Un assistant rend la consultation plus courte, mais le professionnel reprend ensuite la note pendant plusieurs minutes. Le bénéfice apparent change lorsque la mesure couvre l’ensemble de la tâche. Autre possibilité : une note plus complète évite des reprises en aval, sans améliorer immédiatement un critère patient. Ces effets doivent être rapportés séparément.

La transférabilité demande une analyse spécifique : langue, métier, ressources, parcours de soins et accès aux examens. Une adaptation locale ne doit pas être présentée comme une preuve acquise simplement parce que le même modèle est utilisé. Le suivi après déploiement et la gestion des changements appartiennent à l’évaluation.

Niveau	Question à mesurer
Technique	Le système produit-il un résultat fidèle ?
Processus	Le professionnel décide-t-il et travaille-t-il mieux ?
Patient	Un bénéfice ou un risque clinique est-il mesuré ?
Transfert	Le contexte cible est-il suffisamment comparable ?

ANALYSE • ÉTUDES ET PORTÉE

12. La supervision humaine doit être organisée pour fonctionner

La présence d’un professionnel dans le processus ne dit pas encore comment il contrôle le résultat. Peut-il accéder à la source ? Dispose-t-il du temps nécessaire ? Sait-il quelles informations vérifier en priorité ? Peut-il refuser une proposition et signaler une anomalie ?

L’OMS mentionne le biais d’automatisation parmi les risques associés aux grands modèles multimodaux en santé : les utilisateurs peuvent accorder trop de confiance à une sortie du système. Ses recommandations de 2024 appellent une gouvernance adaptée à ces usages. [16]

Je recommande donc une supervision conçue comme une activité de travail, avec un périmètre, des moyens et une responsabilité. Pour un compte rendu, les vérifications prioritaires pourraient porter sur les traitements, les allergies et les décisions prises ; pour une synthèse scientifique, sur les sources, les populations et les critères de jugement. Ces priorités doivent être testées dans chaque usage.

Le gain de temps devrait être mesuré après recherche des sources, relecture et correction. Une tâche achevée plus vite à l’écran peut demander davantage de travail en aval. Inversement, un outil bien conçu peut rendre le contrôle plus facile. La mesure doit permettre de distinguer ces situations.

APPROFONDISSEMENT • MISE EN PRATIQUE

Rendre la supervision faisable

La supervision doit être décrite comme une tâche. Qui reçoit la sortie ? Que doit-il contrôler ? Dispose-t-il de la source, de la compétence et du temps ? À quel moment peut-il interrompre le processus ? Une signature finale ne répond pas à ces questions.

Je propose de distinguer les contrôles rapides sur des éléments critiques et les audits périodiques du fonctionnement. Pour une note clinique, les éléments prioritaires peuvent être les traitements, allergies, décisions et informations de suivi. Cette priorité doit être adaptée à la tâche ; elle ne rend pas acceptable l’abandon des autres contrôles nécessaires.

Le risque augmente lorsque l’outil exécute une action : transmission d’un document, modification d’une prescription ou réservation. **Exemple hypothétique.** Un agent prépare un envoi et l’utilisateur valide sous pression sans relire le destinataire et la pièce jointe. Le problème ne porte plus seulement sur le texte ; il porte sur l’action et sa conséquence. Une confirmation doit présenter ces éléments précisément.

Le biais d’automatisation décrit par l’OMS justifie l’attention au travail humain, mais son existence ne suffit pas à quantifier un préjudice dans chaque contexte. La supervision proposée doit être testée : détecte-t-elle les erreurs, en particulier les erreurs décisives ? Quelle charge ajoute-t-elle ? Quels professionnels ont du mal à l’appliquer ? [16]

Le calcul du gain doit inclure les reprises après signature, les demandes de clarification et la correction des documents transmis. Des utilisateurs peuvent aussi renoncer à l’outil ; leurs motifs doivent être conservés plutôt qu’exclus du bilan comme de simples non-utilisateurs.

Enfin, un dispositif de contrôle ne devrait pas dépendre de l’héroïsme d’une personne déjà surchargée. L’organisation doit attribuer des moyens, prévoir un accès au support et permettre la suspension sans pénaliser celui qui signale une anomalie. L’établissement articule ces dispositions avec ses obligations.

Condition	Élément observable
Compréhension	Usage prévu et limites accessibles.
Accès	Sources et différences visibles.
Temps	Charge de contrôle intégrée au travail.
Pouvoir	Refus, signalement et suspension praticables.

ANALYSE • ÉTUDES ET PORTÉE

13. L’OMS élargit la gouvernance à toute la recherche liée à l’IA

Le rapport **Artificial intelligence-related health research: ethics review and oversight**, publié par l’OMS en juillet 2026 et annoncé en septembre, distingue la recherche de données de santé utilisant l’IA, la recherche conduite avec des outils d’IA et la recherche portant sur ces outils. Sa présentation officielle examine les rôles complémentaires des financeurs, des mécanismes d’accès aux données, des éditeurs et des régulateurs, avec une attention aux inégalités affectant les pays à revenu faible ou intermédiaire. [17]

Ce cadre aide à répartir les responsabilités. Une erreur documentaire appelle un contrôle différent d’un problème de données personnelles, d’une évaluation inadaptée à une population ou d’un risque clinique.

Pour une organisation, je recommande de préciser qui autorise l’usage, qui vérifie le résultat, qui traite les incidents et qui peut suspendre le système. Ces responsabilités doivent rester compréhensibles pour les utilisateurs et pour les personnes concernées par les décisions.

En France, rendre le contrôle praticable

La HAS propose un repère directement utilisable : **A.V.E.C. — Apprendre, Vérifier, Estimer, Communiquer**. Ses Premières clefs d’usage de l’IA générative en santé, publiées en octobre 2025 et actualisées en avril 2026, invitent notamment à comprendre les limites, vérifier les contenus et estimer la pertinence de l’usage. Ce document HAS doit être distingué du projet de guide HAS–CNIL soumis à consultation en 2026. [23]

La gouvernance suppose aussi de tenir les calendriers à jour. Au 5 octobre 2026, la Commission européenne présente les échéances modifiées du règlement sur l’IA : décembre 2027 pour certains systèmes à haut risque et août 2028 pour ceux intégrés à des produits réglementés. L’applicabilité dépend de la qualification du système et de son usage prévu ; toute utilisation médicale d’un chatbot ne relève pas automatiquement de la même catégorie. [24]

APPROFONDISSEMENT • MISE EN PRATIQUE

Organiser la gouvernance autour des décisions

Une gouvernance utile relie les responsabilités aux objets réellement contrôlés. Le responsable du corpus veille à sa provenance et à ses versions ; l’équipe méthodologique examine les analyses ; l’équipe clinique apprécie l’usage ; l’établissement organise les moyens et les incidents. Plusieurs rôles peuvent être tenus par une même personne, à condition de rendre explicites les dépendances.

Le rapport OMS de 2026 élargit la réflexion aux différentes recherches liées à l’IA. La présentation officielle définit le périmètre mobilisé ici. La grille de ce livre organise les responsabilités locales. [17]

En France, le repère A.V.E.C. de la HAS rappelle notamment l’apprentissage, la vérification, l’estimation de la pertinence et la communication. Le passage à une organisation locale demande des documents, des responsables et un processus de réévaluation. Il faut distinguer ce document du projet conjoint HAS–CNIL. [23]

Le calendrier européen doit être consulté à la date d’une décision. Au 5 octobre 2026, la page de la Commission présente une application générale depuis août 2026, avec des exceptions et des échéances ultérieures pour certaines catégories à haut risque. Elle indique notamment décembre 2027 et août 2028 selon les catégories. La décision juridique repose sur le texte applicable, la finalité du système et sa qualification. [24]

Je propose une revue des changements lorsqu’évoluent le modèle, le corpus, l’interface ou les droits d’action. Une nouvelle version peut invalider un résultat de pilote même si le nom commercial reste identique. Le fournisseur doit permettre d’identifier ces changements et leur effet possible sur l’évaluation.

Les contrats peuvent préciser la confidentialité, les usages des données, les notifications de modification et les moyens de récupération. Le contrôle scientifique et la protection des données sont complémentaires : un résultat exact peut provenir d’un traitement non autorisé, et un traitement conforme peut produire une réponse erronée.

Décision	Responsable à nommer localement
Autoriser un usage	Porteur, expert métier et gouvernance concernée.
Changer de version	Responsable produit et référent évaluation.
Traiter un incident	Référent identifié avec pouvoir d’escalade.
Reprendre après arrêt	Instance capable de vérifier la correction.

Qui vérifie quoi ?

Répartition opérationnelle des rôles dans la chaîne de preuve.

Acteur	Provenance	Intégrité	Interprétation	Applicabilité	Correction
Chercheurs	P	P	P	A	P
Auteurs	P	P	P	A	P
Éditeurs	A	P	A	A	P
Établissements	P	P	A	P	P
Utilisateurs cliniques	A	A	P	P	P

P = rôle de pilotage proposé sur les objets de son périmètre. A = contribution ou alerte. Plusieurs acteurs peuvent piloter des objets différents.

La responsabilité doit porter sur une pièce vérifiable.

Nommer qui contrôle la source, les données, la portée, l’usage et les documents déjà diffusés. Préciser les moyens et le pouvoir de suspendre.

Répartition opérationnelle de Lionel Reichardt. Repères : ICMJE, OMS 2026, HAS.

ANALYSE • ÉTUDES ET PORTÉE

14. Ce que je propose de demander à un système d’IA en santé et en recherche

La grille des six préalables

La grille des six préalables structure la décision d’un laboratoire, d’un comité d’éthique ou d’une direction médicale.

- □ **1. Définir la tâche et le bénéfice attendu.** Décrire l’utilisateur, le contexte, le résultat recherché et les usages exclus. Choisir des critères adaptés : temps de documentation, qualité d’une synthèse ou résultat patient.
- □ **2. Rendre le résultat traçable.** Relier les affirmations décisives aux passages, données ou expériences qui les soutiennent. Documenter la version du modèle, le corpus et les transformations importantes.
- □ **3. Vérifier la portée de la preuve.** Contrôler la population, le comparateur, les critères de jugement, l’incertitude et le niveau de validation. Une expérience préclinique ne suffit pas à établir un bénéfice patient.
- □ **4. Organiser un contrôle indépendant.** Désigner les personnes chargées de vérifier les éléments décisifs, leurs moyens et le temps disponible. Prévoir une vérification extérieure à l’appréciation du système qui a produit le résultat.
- □ **5. Mesurer le travail complet et la sécurité.** Intégrer relecture, recherche des sources et correction au calcul du gain. Mesurer les erreurs, leur gravité et leur détection dans le contexte réel d’utilisation.
- □ **6. Prévoir la correction et la suspension.** Désigner un responsable des signalements. Définir comment corriger les documents concernés, informer leurs utilisateurs et interrompre l’usage lorsqu’un problème le justifie.

Pour chaque case, demander un élément concret : protocole, lien vers la source, résultat d’évaluation, responsable nommé ou procédure de signalement. Une case cochée sans moyen de vérifier ce qu’elle recouvre apporte peu de garanties.

Ces exigences donnent une manière concrète de discuter d’un outil avec ses concepteurs. Elles évitent de réduire la confiance à la réputation d’un modèle, à une démonstration impressionnante ou à la qualité de sa rédaction.

L’IA ouvre des possibilités de recherche qui méritent d’être développées. Pour les utiliser en santé, nous devons pouvoir suivre le chemin qui conduit d’une proposition à une expérience, puis d’un résultat à une décision. **C’est ce chemin vérifiable qui donne sa valeur à la preuve.**

APPROFONDISSEMENT • MISE EN PRATIQUE

Passer de la grille à une décision documentée

Les six préalables rendent la décision documentable. Chaque case doit renvoyer à une pièce vérifiable. L’évaluation clinique et réglementaire s’articule avec cette grille selon le rôle de l’outil.

Je propose trois issues : usage autorisé dans un périmètre défini ; expérimentation limitée avec contrôle renforcé ; usage différé ou interrompu jusqu’à résolution d’un point décisif. Les critères d’orientation doivent être fixés selon le rôle réel de l’outil et le risque de l’action, sans construire un score arbitraire additionnant des domaines différents.

Exemple hypothétique. Un assistant prépare une synthèse bibliographique interne. Le laboratoire autorise l’exploration, mais exige une vérification des références et des inférences avant toute diffusion. La même autorisation ne couvre pas l’envoi de recommandations thérapeutiques à des patients. Le changement de destinataire modifie l’usage et appelle une nouvelle décision.

La réévaluation doit avoir une date, un déclencheur et un responsable. Elle peut être déclenchée par une version de modèle, une modification du corpus, un incident ou un changement de population. Un calendrier seul ne remplace pas la réaction à une anomalie importante.

Lors d’un incident, l’organisation doit distinguer le correctif technique, la correction documentaire et l’information des personnes concernées. Corriger le modèle ne supprime pas une note déjà transmise. Inversement, corriger une note n’établit pas que la cause systémique a disparu.

Le but de ce cadre est de rendre la confiance inspectable. Il explicite ce qui est mesuré, ce qui reste incertain et qui peut intervenir. La décision reste proportionnée à la tâche et révisable à partir de nouvelles preuves.

Issue proposée	Ce qui doit être documenté
Autorisation ciblée	Périmètre, contrôles, responsable et réévaluation.
Pilote limité	Protocole, exposition, critères d’arrêt et suivi.
Décision différée	Pièce manquante et condition de reprise.
Suspension	Cause, documents affectés et information à fournir.

La grille des six préalables

Fiche à conserver : les pièces et responsables de la décision.

☐ 1. Tâche et bénéfice

Utilisateur, contexte, objectif et usages exclus.
Élément vérifiable : _____ **Responsable :** _____

☐ 2. Traçabilité

Sources, passages, données, modèle et versions.
Élément vérifiable : _____ **Responsable :** _____

☐ 3. Portée de la preuve

Population, comparateur, critère et incertitude.
Élément vérifiable : _____ **Responsable :** _____

☐ 4. Contrôle indépendant

Personnes, moyens, temps et pièces de vérification.
Élément vérifiable : _____ **Responsable :** _____

☐ 5. Travail complet et sécurité

Relecture, correction, erreurs et gravité.
Élément vérifiable : _____ **Responsable :** _____

☐ 6. Correction et suspension

Signalement, documents affectés et pouvoir d’arrêt.
Élément vérifiable : _____ **Responsable :** _____

Contexte d’usage : _____

Date de réévaluation : _____ Décision : _____

Six préalables de la section 14. Grille de décision de Lionel Reichardt : vérifier les pièces, nommer les responsables, prévoir la réévaluation.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

A. Registre d’audit d’une affirmation

Usage : préparer une synthèse, un support de formation ou une recommandation interne. Fiche de travail.

Commencer par les affirmations qui portent une conclusion ou une décision. Une vérification bibliographique globale ne remplace pas l’examen du passage qui soutient ces affirmations.

Champ	Élément à conserver
Affirmation exacte	Phrase, emplacement et version du document
Rôle dans la décision	Contexte, argument secondaire ou conclusion décisive
Identité de la source	Titre, auteurs, année, DOI ou autre identifiant vérifié
Appui documentaire	Passage, tableau ou figure, page et lien
Protocole	Expérience, enquête, observation, essai ou commentaire
Population et unité	Personnes, articles, références, phrases ou expériences
Résultat	Numérateur, dénominateur, estimation et incertitude
Limite de transfert	Exclusions, contexte et différences avec l’usage prévu
Contrôle	Nom, date et conclusion du vérificateur
Correction	Formulation finale et documents à mettre à jour

Exemple hypothétique. L’affirmation « l’outil évite les erreurs de prescription » s’appuie sur une étude mesurant la fidélité de résumés simulés. La source existe, mais le résultat ne soutient pas cette affirmation. Le registre conduit à reformuler : « le protocole évalue des erreurs de synthèse, sans mesurer un effet sur les prescriptions ». La décision clinique reste à évaluer.

La catégorie « source non retrouvée » doit rester distincte de « fabrication démontrée ». L’échec d’une recherche peut tenir à une mauvaise métadonnée ou à un accès limité. Le contrôleur note ce qu’il a vérifié et ce qu’il ne peut pas trancher.

Lorsque plusieurs affirmations utilisent la même source, elles doivent être examinées séparément. Une étude peut soutenir l’une et pas l’autre. Le registre devient ensuite un outil de correction : il permet de retrouver les supports qui dépendent d’un résultat modifié.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

B. Protocole minimal d’évaluation d’un assistant documentaire

Objectif : identifier où échoue la chaîne recherche–extraction–synthèse. Évaluation locale selon l’usage.

Définir d’abord le domaine, les utilisateurs, les langues et les usages exclus. Constituer un ensemble de questions représentatives, avec des cas difficiles : sources contradictoires, résultat préclinique, absence de différence significative, publication corrigée et population exclue. Les exemples de test doivent être documentés et contrôlés par des experts du domaine.

Dimension	Contrôle proposé
Couverture	Études pertinentes trouvées et manquées
Identité	Concordance des titres, auteurs, dates et identifiants
Extraction	Valeurs, unités, groupes et emplacement dans la source
Soutien	Chaque affirmation décisive correspond-elle au passage cité ?
Interprétation	Association, causalité, préclinique et clinique distinguées
Abstention	Limites, contradictions ou absence de preuve signalées
Actualisation	Versions corrigées et rétractées reconnaissables
Travail complet	Temps de recherche, lecture, vérification et correction

Conserver la version du modèle et du corpus, les requêtes, les documents renvoyés et la réponse. Prévoir une annotation indépendante des erreurs avec des définitions écrites. Rapporter les résultats sur les unités prévues et les désaccords, sans sélectionner seulement les questions réussies.

Un assistant peut être autorisé pour l’exploration tout en restant inadapté à une synthèse de preuve destinée aux soins. La décision doit correspondre au niveau effectivement évalué. Les performances d’OpenScholar et les exigences de traçabilité orientent la construction des critères locaux. [10, 29]

Livrable attendu : rapport daté avec tâches autorisées, limites, exemples d’échecs, moyens de contrôle et déclencheurs de réévaluation. Une moyenne ne doit pas masquer les erreurs rares portant une conclusion décisive.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

C. Évaluer une note clinique : fidélité et utilité

Usage : préparation d’un pilote de documentation clinique. La fiche prépare l’évaluation dans les procédures de l’établissement.

Décrire l’entrée disponible : transcription, dossier, documents annexes et informations structurées. Définir ce que la note doit conserver, sans supposer que toute phrase de la source a la même importance. Les traitements, allergies, décisions et informations de suivi méritent une appréciation métier explicite.

Domaine	Mesure ou observation à prévoir
Ajouts non étayés	Nombre, unité, rôle clinique et exemples
Contradictions	Information source opposée au texte produit
Omissions	Information nécessaire absente ; dénominateur défini
Gravité potentielle	Critères écrits et évaluation compétente
Détection	Erreurs interceptées avant signature ou transmission
Correction	Temps, difficulté, reprises et charge cognitive
Résultat patient	À mesurer dans un protocole adapté, pas à déduire du texte
Variabilité	Professionnels, langues, tâches et situations difficiles

Les notes multiples issues d’une même consultation ne doivent pas être traitées comme des patients indépendants. Une analyse doit tenir compte de la structure du corpus. L’expérience d’Asgari fournit des résultats situés et rappelle l’importance de distinguer fréquence par phrase et gravité parmi les erreurs. [9]

Exemple hypothétique. Une allergie est présente dans le dossier mais omise dans la note. Le texte ne contient pas de phrase fausse ; il reste incomplet pour une décision de prescription. Le contrôle de l’exhaustivité utile détecte ce que le seul comptage des inventions pourrait manquer.

Avant le pilote, définir les critères d’arrêt et le circuit de signalement. Après le pilote, conserver les cas où l’utilisateur renonce à l’assistant ou doit reprendre entièrement la note. Ces situations appartiennent au bilan du travail réel.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

D. Journal des actions d’un agent

Usage : recherche automatisée, programmation et préparation de transmissions. Journal de traçabilité ; les données personnelles et secrets sont protégés selon le contexte.

Champ	Contenu à enregistrer
Demande initiale	Objectif, auteur de la demande et périmètre autorisé
État de départ	Versions, fichiers et données disponibles
Action proposée	Lecture, modification, exécution ou transmission
Objet exact	Fichier, table, script, destinataire ou autre cible
Autorisation	Règle applicable et personne qui approuve si nécessaire
Action réalisée	Opération effective et horodatage
Résultat	Sortie, erreur, coût et effet observable
Intervention humaine	Modification, refus ou reprise
Réversibilité	Sauvegarde, retour et condition d’arrêt
Clôture	Vérification finale et limites restantes

La trace distingue « proposé », « approuvé » et « exécuté ». Une déclaration rétrospective de l’agent ne suffit pas à établir qu’une action a eu lieu. Pour des modifications, conserver les différences ; pour une analyse, garder les paramètres et sorties ; pour une transmission, garder le contenu final et le destinataire.

Exemple hypothétique. Un agent met à jour un tableau de résultats puis prépare l’envoi d’un rapport. La validation du calcul et l’autorisation d’envoyer sont deux contrôles. L’utilisateur doit pouvoir voir les modifications et l’objet de l’envoi, sans reconstruire mentalement toute la séquence.

Prévoir des droits proportionnés. Un agent chargé d’explorer la littérature peut avoir accès en lecture à un corpus sans disposer d’un droit de publication. Un agent chargé de programmer peut exécuter dans un environnement isolé sans pouvoir modifier les données originales.

La supervision devient une activité de gouvernance des actions. Les journaux doivent rester exploitables par une autre personne que celle qui a lancé l’outil et par une équipe capable d’enquêter après un incident.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

E. Une fiche de signalement et de correction

Objectif : éviter qu’un correctif technique laisse circuler des documents erronés. Le circuit de correction s’articule avec celui de l’établissement et de la revue.

Champ	Information utile
Signalement	Date, personne, canal et description observable
Usage	Tâche, système, version, contexte et destinataires
Objet affecté	Réponse, note, support, analyse ou action
Éléments conservés	Entrée, sortie, source et traces pertinentes
Qualification initiale	Erreur, omission, interprétation ou anomalie à confirmer
Conséquence	Détectée, transmise, décision affectée ou événement observé
Mesure immédiate	Limitation, arrêt, correction ou vérification renforcée
Documents dérivés	Versions, transmissions et synthèses à rechercher
Analyse de cause	Faits établis, hypothèses et inconnues
Reprise	Conditions vérifiées et responsable de décision

L’investigation doit préserver les données nécessaires sans diffuser inutilement des informations de santé. Un incident peut provenir du corpus, du modèle, de l’interface, d’une instruction ou du processus humain. La catégorie ne doit pas être fixée avant l’examen des traces.

Exemple hypothétique. Une citation fictive a été utilisée dans un support de formation. L’organisation corrige la diapositive, signale la modification aux destinataires et vérifie si la conclusion dépendait de cette source. Changer seulement le fichier maître ne suffit pas si des versions antérieures circulent.

Le bilan distingue fréquence d’incidents, exposition, gravité et détection. Une hausse de signalements peut traduire un meilleur dispositif de déclaration plutôt qu’une dégradation du système. Il faut interpréter la série avec son contexte.

La reprise après suspension doit reposer sur un élément vérifiable : correction du corpus, modification testée, contrôle supplémentaire ou restriction du périmètre. L’absence de nouvel incident pendant une courte période ne prouve pas que la cause a disparu.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

F. Feuille de route : cadrer, tester, décider

Trois séquences de décision. Les durées et effectifs dépendent de l’usage ; aucun calendrier universel de validation n’est imposé.

Séquence	Travail à réaliser	Livrable de décision
Cadrer	Décrire tâche, population, bénéfice, risques, données et usages exclus	Fiche d’usage et responsabilités
Tester	Évaluer corpus, erreurs, contrôle humain et temps complet	Rapport de pilote avec limites
Décider	Examiner bénéfice, sécurité, moyens, corrections et changements	Autorisation ciblée, poursuite du pilote ou arrêt

Cadrer. Partir d’une difficulté réelle du travail. Définir ce qui serait une amélioration utile, ce qui pourrait être dégradé et quelles actions le système pourra effectuer. Identifier les exigences scientifiques, organisationnelles, de confidentialité et réglementaires applicables. Désigner les personnes capables d’évaluer chaque domaine.

Tester. Construire un comparateur pertinent et conserver les tâches échouées. Mesurer la production et la validation séparément, puis ensemble. Observer les reprises, les omissions et les erreurs qui portent une décision. Le contrôle humain doit être évalué comme une composante du système, avec sa charge et ses limites.

Décider. Ne pas transformer un pilote réussi en autorisation générale. Documenter le périmètre, les pièces disponibles, les incertitudes, les changements qui imposent une nouvelle évaluation et le pouvoir de suspendre. Une décision différée doit préciser l’élément manquant plutôt que rester une réserve vague.

Les utilisateurs concernés doivent comprendre les contrôles attendus et disposer de moyens pour les accomplir. L’information d’un patient ou d’un participant à une recherche doit être adaptée à l’usage et aux exigences applicables ; une formule générale de consentement ne remplace pas cet examen.

Point de contrôle final : la confiance est-elle attachée à une tâche évaluée et à une version identifiable ? Peut-on retrouver ce qui soutient la décision, corriger les documents et interrompre une action ? Si ces réponses sont absentes, le projet demande encore du travail de gouvernance.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

G. Formuler les résultats avec précision

Cette page rassemble les limites qui doivent accompagner la diffusion des chiffres. Elle fait partie du raisonnement scientifique ; elle ne vise pas à minimiser les apports de l’IA.

Formulation trop large	Formulation soutenue par le corpus
« Les publications explosent à cause de ChatGPT »	La croissance documentaire est observée avant les LLM ; l’effet causal exige un protocole spécifique.
« L’IA fait gagner sept heures à chaque chercheur »	6,9 heures en moyenne sont déclarées dans une enquête située de 637 chercheurs. [18]
« Les soumissions mondiales ont augmenté de 42 % »	+42 % est rapporté dans une seule revue. [19]
« 44 % des patients subissent une erreur »	84 des 191 hallucinations sont classées majeures dans une expérience simulée. [9]
« Des milliers de références prouvent une fraude par IA »	L’audit détecte des fabrications sans attribution causale à l’IA. [7]
« L’agent a publié à ICLR »	Un manuscrit atteint un seuil d’acceptation dans un atelier ; les soumissions sont retirées selon le protocole. [3]
« 335 articles ont été fabriqués par IA »	Il s’agit d’un corpus de rétractations d’articles portant sur l’IA. [21]
« Une source réelle suffit à valider une décision »	Il faut examiner le soutien à l’affirmation et l’applicabilité.
« La littérature a subi un model collapse »	Les expériences d’entraînement récursif ne mesurent pas un effondrement global de la littérature. [22]

Un chiffre sans unité peut être exact et trompeur. Un résultat publié peut être réel et discuté. Une absence de différence peut être compatible avec plusieurs effets possibles. La rigueur consiste à conserver ces limites au plus près de l’affirmation, pas dans une réserve générale que le lecteur devrait retrouver ailleurs.

Pour une présentation publique, préférer moins de chiffres et davantage de précision sur ceux qui restent. La fiche visuelle n’exonère pas de fournir le protocole et la source accessibles au lecteur.

ANNEXE PRATIQUE • OUTIL DE TRAVAIL

H. Carte des preuves et des décisions

Les références ne jouent pas toutes le même rôle. Cette carte oriente la lecture sans classer les études sur une échelle commune de performance.

Question	Sources principales	Nature de la pièce
Que produit une IA scientifique spécialisée ?	AlphaFold 3 [1], Co-Scientist [2]	Prédiction et validations précliniques
Jusqu’où automatiser une recherche ?	The AI Scientist [3]	Démonstration encadrée en apprentissage automatique
L’usage augmente-t-il la production ?	Kusumegi [4], Renault et al. [5]	Analyse et contre-analyse de l’identification
Que devient la diversité scientifique ?	Hao et al. [6], Messeri et Crockett [32]	Observation et analyse conceptuelle
Quelles références sont fabriquées ?	Topaz [7], Walters et Wilder [31]	Audit de littérature et expérience historique
Quel risque dans les notes ?	Asgari [9]	Expérience de synthèse simulée
Que change la recherche documentaire ?	OpenScholar [10]	Évaluation de recherche et synthèse
Quel bénéfice pour les patients ?	Agweyu [15]	Essai pragmatique randomisé en grappes
Comment rapporter et gouverner ?	PRISMA [29], DECIDE-AI [30], OMS [16, 17], HAS [23]	Recommandations de rapport et de gouvernance
Quel contexte de volume ?	Hanson [27], NSF/NCSES [28]	Bibliométrie et série statistique institutionnelle

Un essai répond à une question située ; un cadre institutionnel aide à organiser des responsabilités ; une enquête décrit des déclarations ; une analyse conceptuelle formule des mécanismes à examiner. Le passage de l’un à l’autre exige une justification.

Pour une décision locale, commencer par le type d’usage et la conséquence possible. Chercher ensuite la pièce qui peut réellement répondre à cette question. Un score de génération de texte, aussi élevé soit-il, n’est pas la pièce adaptée à une demande de bénéfice patient.

Cette carte facilite également la mise à jour : lorsque paraît un nouveau travail, préciser quelle question il modifie, ce qu’il mesure et ce qu’il laisse inchangé. L’ajout d’une référence ne constitue pas, à lui seul, une amélioration de la preuve.

BIBLIOGRAPHIE • 33 RÉFÉRENCES

Sources et repères de lecture

Les références sont accompagnées de leur protocole, de leur niveau de validation et de leurs limites d’accès. Le contrat méthodologique figure page 4.

1. **Abramson et al.** Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024. DOI 10.1038/s41586-024-07487-w. Prédiction de structures, comparaisons expérimentales de référence et limites du modèle.

2. **Co-Scientist.** Accelerating scientific discovery with Co-Scientist. Nature, 2026. DOI 10.1038/s41586-026-10644-y. Génération d’hypothèses et validations biomédicales précliniques.

3. **Lu et al.** Towards end-to-end automation of AI research. Nature, 2026. DOI 10.1038/s41586-026-10265-5. Résultats spécifiques en apprentissage automatique ; texte intégral consulté. Trois manuscrits sélectionnés par des humains soumis à l’atelier ICBINB ; un atteint le seuil d’acceptation, puis retrait selon le protocole. Ne pas confondre atelier et conférence principale.

4. **Kusumegi et al.** Scientific production in the era of large language models. Science, 2025. DOI 10.1126/science.adw3000. Version auteur. Production et indicateurs indirects de qualité ; interprétation causale contestée dans la référence suivante.

5. **Renault, Bergeaud et Bosquet.** Scientific production in the era of large language models: Outcome-triggered treatment timing and spurious event-study dynamics. PNAS, 2026. DOI 10.1073/pnas.2618638123. Version arXiv consultée. Critique de la datation de l’adoption, simulations et placebos ; publication finale identifiée.

6. **Hao et al.** Artificial intelligence tools expand scientists’ impact but contract science’s focus. Nature, 2026. DOI 10.1038/s41586-025-09922-y. Analyse observationnelle de la production et de l’exploration collectives, portant sur des outils d’IA au sens large.

7. **Topaz et al.** Fabricated citations: an audit across 2·5 million biomedical papers. The Lancet, Correspondence, 9 mai 2026, 407, 1779–1781. Texte original hébergé par l’auteur. Texte principal consulté ; annexe non consultée. Fabrications dans un corpus publié, sans attribution causale à l’IA.

8. **Peoples et al.** When machines misread science. npj Digital Medicine, 2026. DOI 10.1038/s41746-026-03160-w. Commentaire et propositions sur l’interprétation des publications, non dispositif validé.

9. **Asgari et al.** A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine, 2025. DOI 10.1038/s41746-025-01670-7. Consultations simulées, évaluations répétées ; gravité potentielle et affiliation industrielle à considérer.

10. **Asai et al.** Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026. DOI 10.1038/s41586-025-10072-4. OpenScholar et évaluation multidisciplinaire de recherche et de synthèse.

11. **Sarkar.** Why AI can’t be trusted to write scientific reviews. Nature, World View, 26 mai 2026. DOI 10.1038/d41586-026-01616-3. Introduction accessible consultée ; reste sous abonnement. Témoignage expert de la rédactrice en chef de Cochrane.

12. **ICMJE.** AI use by authors et Defining the Role of Authors and Contributors. Responsabilité, qualité d’auteur et déclaration. Les instructions de la revue restent à vérifier pour chaque soumission.

13. **Elsevier.** Generative AI policies for journals. Politiques relatives aux auteurs, éditeurs et relecteurs ; exigences dépendant du rôle et de l’usage.

14. **Thelwall.** Research quality evaluation by AI in the era of Large Language Models: Advantages, disadvantages, and systemic effects. arXiv:2506.07748v1, 9 juin 2025. Prépublication, PDF fourni consulté ; aucune publication finale confirmée dans cet audit.

15. **Agweyu et al.** Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial. Nature Medicine, 26 juin 2026. DOI 10.1038/s41591-026-04503-6. Essai dans des structures de soins primaires au Kenya, avec critère patient à 14 jours.

16. **OMS.** Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. 2024. Annonce officielle du 18 janvier 2024 ; fiche du rapport. Gouvernance des modèles multimodaux et risques, dont le biais d’automatisation.

17. **OMS.** Artificial intelligence-related health research: ethics review and oversight. 21 juillet 2026, ISBN 978-92-4-012407-3. Fiche officielle ; communiqué du 21 septembre 2026. Fiche et communiqué consultés ; rapport intégral non récupéré.

18. **Codreanu et al.** AI in Science: Early Insights. Rapport de recherche, septembre 2026. PDF original. Google, Google DeepMind, MIT FutureTech et autres affiliations ; enquête déclarative, données d’usage et inventaire de modèles spécialisés. Ne pas assimiler ce rapport à un essai causal.

19. **Gartenberg, Hasan, Murray et Pierce.** More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review. Organization Science, 27 avril 2026. DOI 10.1287/orsc.2026.ed.v37.n3. Analyse d’une revue, détection statistique agrégée et indicateurs rédactionnels ; portée distincte de la validité des résultats.

20. **arXiv.** Attention Authors: Updated Practice for Review Articles and Position Papers in arXiv CS Category. 31 octobre 2025. Annonce officielle. Périmètre informatique et types de soumissions précisés.

21. **Sridharan et Sivaramakrishnan.** Artificial intelligence in the retraction spotlight: trends, causes and consequences of withdrawn AI literature through a systematic bibliometric review. Frontiers in Research Metrics and Analytics, 20 janvier 2026. DOI 10.3389/frma.2025.1737168. Corpus d’articles portant sur l’IA ; recherche dans Scopus en avril 2025. L’année du DOI et du volume est 2025, la publication est datée de janvier 2026.

22. **Shumailov et al.** AI models collapse when trained on recursively generated data. Nature, 24 juillet 2024. DOI 10.1038/s41586-024-07566-y. Mécanisme et expériences d’entraînement récursif ; aucune mesure de l’effondrement d’un corpus scientifique global.

23. **HAS.** Premières clefs d’usage de l’IA générative en santé. 30 octobre 2025, actualisation du 15 avril 2026. Document officiel. Repère A.V.E.C. ; document distinct du projet conjoint HAS–CNIL.

24. **Commission européenne.** AI Act — cadre réglementaire et calendrier actualisé. État consulté au 5 octobre 2026 ; modifications du calendrier et distinction des catégories à haut risque. La qualification juridique dépend du produit et de son usage.

25. **Cabanac, Labbé et Magazinov.** Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. arXiv:2107.06751, juillet 2021. Prépublication sur des reformulations suspectes antérieures à la diffusion de ChatGPT.

26. **Sharma et al.** Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, soumis en 2023, version révisée en mai 2025. Notice et résumé consultés ; expériences sur des assistants ajustés avec des préférences humaines, sans mesure clinique.

27. **Hanson, Gómez Barreiro, Crosetto et Brockington.** The strain on scientific publishing. Quantitative Science Studies, 2024, 5(4), 823–843. DOI 10.1162/qss_a_00327. Version auteur. Analyse bibliométrique de croissance et de tensions éditoriales ; périmètre distinct de la série NSF.

28. **National Science Board / NCSES, NSF.** The State of U.S. Science and Engineering 2026, figure 29, publications mondiales en sciences et ingénierie, 2014–2024. Série et tableau officiels. Rapport publié en mai 2026 ; données Scopus extraites en août 2025. Articles de revues et communications de conférences sélectionnés. Les variations du livre sont calculées à partir du tableau.

29. **Page et al.** The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 2021, 372:n71. DOI 10.1136/bmj.n71. Recommandation de compte rendu ; ne certifie pas la qualité méthodologique d’une revue.

30. **Vasey et al.** Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 2022, 28, 924–933. DOI 10.1038/s41591-022-01772-9. Cadre de compte rendu des évaluations cliniques précoces.

31. **Walters et Wilder.** Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 2023, 13:14045. DOI 10.1038/s41598-023-41032-5. Expérience historique de génération bibliographique avec des versions précises ; pas un taux actuel de tous les assistants.

32. **Messeri et Crockett.** Artificial intelligence and illusions of understanding in scientific research. Nature, 2024, 627, 49–58. DOI 10.1038/s41586-024-07146-0. Analyse conceptuelle des promesses, des illusions et des risques de monocultures scientifiques.

33. **NIST.** Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4, novembre 2024. DOI 10.6028/NIST.AI.100-4. Rapport officiel. Marquage, métadonnées, détection et évaluation de la robustesse ; limites de suppression et de falsification du watermarking.

Lectures complémentaires du corpus transmis

- VoxUkraine, Fake Citations, 10 juillet 2026 : porte d’entrée vers l’audit original du Lancet.
- Berkeley Haas, 27 janvier 2026 : présentation institutionnelle des travaux de Kusumegi et ses collègues, à lire avec la contre-analyse.
- Krammer, The Conversation, republication attribuée du 7 mai 2026 : analyse prospective de l’automatisation scientifique ; conserver la distinction entre l’évaluation du système et ses productions.
- **Khlaif et al.**, The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation, JMIR Medical Education, 2023, DOI 10.2196/47049 : protocole exploratoire historique, quatre articles générés et 23 relecteurs ; notice et extraits indexés consultés.
- The Guardian, 13 juillet 2025 : contexte de la pression éditoriale.

PHARMAGEEK • CONFÉRENCES ET ATELIERS

Travaillons ensemble

Faire de la preuve un sujet de travail, de formation et de décision.

Chercheurs, professionnels de santé, équipes éditoriales, facultés, institutions : vos besoins ne se résument pas à découvrir un nouvel outil. Il s’agit de comprendre ce qu’il permet, ce qu’il exige de vérifier et ce qu’il change dans l’organisation.

Conférences et débats

Mettre en perspective les avancées, les limites et les usages de l’IA en santé. Donner des repères documentés pour discuter la promesse, la preuve et les effets sur les pratiques.

Ateliers et formation

Travailler sur des cas adaptés à vos métiers : vérifier une référence, lire un dénominateur, distinguer préclinique et clinique, auditer une synthèse ou préparer une grille d’usage.

Échanges avec vos équipes

Construire les bonnes questions avant une expérimentation : bénéfice recherché, contrôle humain, responsabilités, évaluation et correction. Les fiches de ce livre peuvent servir de point de départ à cette discussion.

Préparer. Animer. Prolonger.

Relier l’intervention à votre contexte, faciliter les échanges et prolonger la réflexion par des contenus et ressources. Les deux pages suivantes présentent l’univers Pharmageek et les formats d’intervention.

Lionel Reichardt — Pharmageek

lionel@reichardtonline.com

LinkedIn : Lionel Reichardt

Quel usage de l’IA mérite aujourd’hui une meilleure vérification dans votre organisation ?

Pharmageek
IAtrogénique × IAtus
Lionel Reichardt

LIONEL REICHARDT / PHARMAGEEK

Préparer. Animer. Prolonger.



Un partenaire pour faire de votre événement une conversation utile et un contenu qui dure.

Je prépare les échanges avec vos intervenants, construis le fil rouge et anime les conférences, tables rondes et débats. Mon travail consiste à rendre les sujets de santé numérique, d'innovation et d'IA accessibles tout en conservant leur exigence, et à relier les expertises aux questions de votre public.

Avant

Clarifier les objectifs ; travailler avec les intervenants ; construire les questions et transitions.

Pendant

Faire dialoguer les expertises ; solliciter la salle ; faire émerger les convergences et désaccords.

Après

Prolonger les échanges par des articles, newsletters, vidéos et synthèses adaptés au projet.

Un écosystème de veille et de diffusion

Canal	Repères d'audience et de contenu
Scoop.it	60 000 partages · plus de 2,4 millions de vues
LinkedIn	Plus de 29 000 contacts
IAtrogénique × IAtus	Plus de 5 100 abonnés
Morning Routine	Plus de 3 200 abonnés
YouTube	Plus de 220 vidéos · plus de 740 abonnés

Chiffres communiqués par Lionel Reichardt pour cette édition, octobre 2026.
Les audiences sont propres à chaque canal et peuvent se recouper.

Pharmageek
IAtrogénique × IAtus
Lionel Reichardt

CONFÉRENCES / ANIMATION / RELAIS DE COMMUNICATION

Plus de 100 clients

Quelques références parmi les plus de 100 clients que j'ai accompagnés.

Institutions, industriels, sociétés savantes et médias : des interventions et des collaborations adaptées aux objectifs, au public et au format de chaque événement.

Agence du Numérique en Santé	GIE SESAM-Vitale	SofIA-Santé
Mylan	Cegedim Santé	Groupe Profession Santé
Société Française de Cardiologie	Sanofi	Johnson & Johnson

Références communiquées par Lionel Reichardt. Sélection non exhaustive. Logos : marques de leurs titulaires respectifs ; crédit Johnson & Johnson.

Construisons votre prochain événement.

Vous souhaitez m'avoir comme conférencier, animateur ou partenaire pour votre congrès, votre journée professionnelle ou vos tables rondes ? Échangeons sur vos objectifs, la préparation des interventions et les relais de communication qui prolongeront l'événement.

Me contacter sur LinkedIn
lionel@reichardtonline.com

IAtus × IAtrogénique

Produire plus.
Vérifier mieux.
Décider avec preuve.

Quand l’IA fait science

L’IA peut accélérer un texte, une recherche et une hypothèse. La valeur scientifique dépend encore de ce qui permet de vérifier, de reproduire et d’interpréter le résultat.

Ce livre blanc examine les avancées, les fabrications bibliographiques, les erreurs de synthèse et les responsabilités dans la chaîne de preuve en santé.

33 références • 7 infographies • 8 annexes pratiques

Études, protocoles, limites et recommandations éditoriales distingués.

Lionel Reichardt — Pharmageek

lionel@reichardtonline.com

IAtus × IAtrogénique • Octobre 2026