

GUIDE 04

Une réponse convaincante n'est pas une preuve

Niveau 3 — apprendre à distinguer fluidité, plausibilité, preuve et validation.

ÉCHELLE DE CONFIANCE



Le ton assuré ne fait pas monter d'un cran.

*Au-dessus du seuil, on peut auditer.
En dessous, on ne peut que croire.*

Emplacement réservé à l'infographie de la collection (format portrait). Le schéma ci-dessus tient lieu de repère de mise en page.

Pourquoi notre intuition se trompe

Les modèles génératifs sont conçus pour produire une suite de texte cohérente et utile. Cette qualité de formulation crée une impression de certitude qui n'est pas équivalente à une preuve.

Le NIST nomme **confabulation** « la production de contenu erroné ou faux, énoncé avec assurance — familièrement appelé hallucination — par lequel les utilisateurs peuvent être induits en erreur ». [2]

Le même document précise le mécanisme du risque : les utilisateurs croient le contenu faux *en raison même de l'assurance de la réponse*, puis agissent sur cette base ou la relaient. [2] Ce n'est donc pas un défaut de vigilance : c'est la forme de la réponse qui produit l'effet.

Quatre questions de contrôle

- 1 **Source**
Quel document ou quelle donnée identifiable ?
- 2 **Passage**
Quel élément précis soutient ce point ?
- 3 **Date et périmètre**
Encore valable, et applicable à ce cas ?
- 4 **Statut**
Fait, inférence ou hypothèse ?

Le déplacement à opérer. Passer de « est-ce que ça sonne juste ? » à « qu'est-ce qui permet de vérifier cette affirmation ? ». La première question porte sur le texte ; seule la seconde porte sur le monde.

À RETENIR

Sans preuve, la confiance n'est qu'une impression — et l'assurance du ton n'est pas un indice de qualité.

VERSION WEB PERMANENTE

consulting.amplify-belgium.com/ressources/une-reponse-convaincante-n-est-pas-une-preuve

GUIDE 04 · COLLECTION KNOWLEDGE CLAUDE

Vérifié le 28 août 2026 · 8 pages

GUIDE 04 · 01

Quatre dimensions à ne pas confondre

Une réponse peut être excellente sur une dimension et faible sur une autre. C'est précisément ce qui rend l'évaluation «au feeling» peu fiable.

DIMENSION	QUESTION DE CONTRÔLE	CE QUI COMPTE
Cohérence	Le raisonnement se tient-il ?	Absence de contradiction interne, structure logique, termes employés de façon stable.
Exactitude	Les faits sont-ils corrects ?	Concordance avec des sources fiables et actuelles ; absence de données inventées.
Traçabilité	Peut-on retrouver d'où vient chaque point important ?	Source identifiable, passage ou donnée précise, date et périmètre.
Calibration	La réponse dit-elle correctement ce qui est certain, probable ou inconnu ?	Incertitudes visibles, hypothèses séparées des faits, autorisation explicite de dire «je ne sais pas». ^[1]

La combinaison dangereuse. Une réponse très cohérente et très mal calibrée est la plus trompeuse de toutes : elle se lit bien, ne se contredit jamais, et ne signale rien. La cohérence est la dimension que le modèle réussit le plus facilement — et la seule que l'œil détecte sans effort.

Le piège des citations inventées

Une référence écrite par le modèle n'est pas automatiquement une référence valide. Le NIST note que les systèmes génératifs peuvent produire des citations ou des justifications fabriquées, qui renforcent artificiellement la confiance. ^[2]

La stratégie robuste s'inverse : partir de sources réelles que vous fournissez, puis demander quels passages soutiennent chaque affirmation — plutôt que de demander des sources à posteriori.

Le piège de l'accord

Un exercice pilote d'évaluation croisée mené par Anthropic et OpenAI (août 2025) a observé des comportements de complaisance dans l'ensemble des modèles testés des deux laboratoires, particulièrement marqués sur les modèles généralistes haut de gamme. ^[3]

Cela ne signifie pas que Claude acquiesce systématiquement. Cela signifie qu'un «oui, votre analyse est correcte» ne doit jamais servir de validation factuelle : c'est un énoncé de niveau 1 sur l'échelle de confiance.

À RETENIR

Demander «es-tu sûr ?» ne teste rien. Demander «sur quel passage t'appuies-tu ?» teste quelque chose.

GUIDE 04 · 02

Faire produire une réponse vérifiable

Le bon usage consiste à organiser la réponse autour de preuves, d'incertitudes et de statuts explicites — dès la demande, pas après coup.

LEVIER 1

Autoriser l'incertitude

Si l'information n'est pas établie par les sources fournies, dis « non déterminé » et indique ce qui manque. N'invente pas de réponse pour combler le vide.

Anthropic recommande de donner explicitement à Claude la permission d'admettre son incertitude, et note que cette technique simple peut réduire fortement les informations fausses. ^[1]

LEVIER 2

Ancrer dans les passages

Avant de conclure, extrais mot à mot les passages qui soutiennent chaque point important. Puis fais la synthèse uniquement à partir de ces passages.

Pour les documents longs, la documentation recommande d'extraire les citations *avant* de traiter la tâche : cela ancre la réponse dans le texte réel et réduit les hallucinations. ^{[1][4]}

LEVIER 3

Séparer fait et inférence

Pour chaque conclusion, indique son statut : fait explicite / inférence / hypothèse / information manquante.

Cette séparation empêche une déduction raisonnable d'être présentée comme un fait fourni par le document. C'est le point où se logent la plupart des erreurs de bonne foi.

LEVIER 4

Vérifier date et périmètre

Pour les chiffres, règles ou engagements, indique la source, la date et le périmètre. Signale toute incompatibilité entre documents.

Une donnée exacte mais périmée, ou exacte dans un autre périmètre, est aussi trompeuse qu'une donnée fausse — et beaucoup plus difficile à repérer.

À RETENIR

Ne demandez pas seulement « donne les sources ». Donnez des sources, puis exigez un lien explicite entre chaque affirmation et sa preuve.

GUIDE 04 · 03

Vérifier une affirmation importante

Exemple : une conclusion contractuelle produite à partir d'un dossier interne.

AFFIRMATION	SOURCE	PASSAGE	STATUT
« Le contrat prévoit un préavis de 60 jours. »	Contrat signé — v3	Article 12.2 — préavis	Étayé — à valider
« Nous pouvons résilier sans pénalité. »	Contrat v3 + note juridique interne	Article 14, lu avec l'article 12	Inféré — validation juriste requise
« Le fournisseur s'engage à 99,9 % de disponibilité. »	E-mail commercial de mars	Aucun article contractuel correspondant	Affirmé — inutilisable en l'état

La méthode en cinq étapes

1 Isoler l'affirmation

Écrire la phrase exacte à vérifier. Éviter les formulations vagues qui mélangent plusieurs idées : une affirmation composite ne se vérifie pas, elle se discute.

2 Identifier la source de vérité

Quel document, système ou personne fait foi ? Un résumé secondaire n'a pas le même statut qu'un contrat signé.

3 Retrouver le passage

Demander le passage exact, pas une simple référence de document — puis vérifier que le passage dit réellement ce que la réponse affirme.

4 Contrôler date et portée

Version du document, période concernée, exceptions, conditions, pays, entité juridique.

5 Donner un statut

Validé / étayé / inféré / non déterminé. Si le point est à fort enjeu, faire valider par la personne compétente.

CE QUI NE CONSTITUE PAS UNE PREUVE

« Claude a donné la même réponse deux fois », « il a l'air sûr de lui », « il est d'accord avec mon intuition », « il a cité un nom de document ». Aucun de ces signaux n'est externe à la formulation du modèle — et c'est précisément ce qu'exige une preuve.

Nuance utile sur le premier : la répétition d'un même prompt est un test valable, mais dans l'autre sens. La documentation la présente comme un signal d'alerte quand les réponses *divergent* — une constance, elle, ne prouve rien.^[1]

À RETENIR

Vérifier, ce n'est pas relire une réponse. C'est remonter d'une phrase à un document, puis d'un document à une date.

GUIDE 04 · 04

Les techniques documentées, et ce qu'elles coûtent

Anthropic publie une liste de techniques de réduction des hallucinations. Elles n'ont ni la même efficacité, ni le même coût d'usage. ^[1]

TECHNIQUE	CE QU'ELLE FAIT	QUAND L'EMPLOYER
Permission d'ignorer	Autoriser explicitement Claude à admettre son incertitude — technique simple, effet important sur les informations fausses.	Toujours. C'est le réglage au meilleur rapport effet/effort.
Citations d'abord	Faire extraire les passages mot à mot avant de traiter la tâche, pour ancrer la réponse dans le texte réel.	Documents longs — la documentation cite le seuil indicatif de 20 000 tokens.
Vérification après coup	Faire chercher, pour chaque affirmation, une citation qui la soutient ; si elle est introuvable, l'affirmation doit être retirée.	Livrables destinés à un tiers. C'est la technique qui produit une rétractation, pas seulement un doute.
Raisonnement explicite	Faire exposer le raisonnement étape par étape avant la conclusion, ce qui révèle les logiques ou hypothèses défailtantes.	Analyses et arbitrages, où l'erreur porte sur l'enchaînement plus que sur le fait.
Répétition et comparaison	Relancer le même prompt plusieurs fois : des divergences entre sorties signalent un risque d'hallucination.	Points isolés à fort enjeu. Coûteux, à réserver aux affirmations décisives.
Restriction aux sources	Interdire l'usage de la connaissance générale du modèle et limiter la réponse aux documents fournis.	Travail documentaire, dossiers internes, tout contexte où seule votre matière fait foi.

FORMULATION TYPE — VÉRIFICATION AVEC RÉTRACTATION

Reprends chaque affirmation importante.
Pour chacune, trouve dans les documents fournis la citation qui la soutient.
Si tu n'en trouves pas, retire l'affirmation et dis-le explicitement.

La limite annoncée par la source. La documentation est explicite : ces techniques réduisent significativement les hallucinations, mais ne les éliminent pas. Elle recommande de toujours valider les informations critiques, en particulier pour les décisions à fort enjeu. ^[1]

Autrement dit, aucune de ces techniques ne remplace l'étape humaine : elles déterminent seulement où cette étape doit se concentrer.

À RETENIR

Ces techniques ne rendent pas la réponse vraie. Elles la rendent contestable — ce qui est exactement ce dont vous avez besoin.

GUIDE 04 · 05

Citations, recherche et contrôle de source

Le niveau avancé consiste à rendre la provenance techniquement traçable — sans confondre traçabilité et vérité.

Dans claude.ai

Pour un usage bureautique, demandez le nom du document, le passage ou la citation directe, puis comparez-les à la source. Pour une recherche web, demandez la date, la source primaire si possible, et la distinction entre ce qui est confirmé et ce qui reste incertain.

Dans la Claude Platform / API

La fonctionnalité Citations rattache les réponses à des passages exacts des documents fournis. Anthropic indique que ces pointeurs sont extraits directement des documents — ils sont donc garantis de pointer vers le document réel — et que, dans ses évaluations, la fonctionnalité cite les passages les plus pertinents nettement plus souvent qu'une approche uniquement par prompt. ^[5]

Ce que la fonctionnalité garantit — et ce qu'elle ne garantit pas

ELLE AIDE À GARANTIR	ELLE NE GARANTIT PAS
Que le pointeur renvoie à un passage réel du document fourni. ^[5]	Que le document source est vrai, à jour ou applicable à votre situation.
Que la provenance d'une affirmation est facilement vérifiable.	Que l'interprétation du passage est juridiquement, financièrement ou techniquement correcte.
Que le lecteur peut auditer la chaîne source → affirmation.	Que toutes les affirmations importantes ont été citées, ni que rien n'a été omis.

Deux limites techniques à connaître. Seules les citations de texte sont prises en charge : les citations d'images, y compris depuis un PDF, ne le sont pas. Et la fonctionnalité n'est pas combinable avec les sorties structurées. ^[5]

Pour les recherches complexes

Les bonnes pratiques de prompting recommandent de définir les critères de réussite, de recouper l'information sur plusieurs sources et, sur les sujets complexes, de maintenir plusieurs hypothèses concurrentes avec des niveaux de confiance et des auto-critiques régulières. ^[4] Cela transforme la recherche en enquête, plutôt qu'en production d'une réponse séduisante.

À RETENIR

La traçabilité permet d'auditer une réponse. Elle ne dispense jamais d'évaluer la qualité de la source et la justesse de l'interprétation.

GUIDE 04 · 06

Ce que la vérification ne résout pas seule

La preuve est nécessaire. Le jugement sur la qualité, la pertinence et les conséquences reste un travail humain.

Une source peut être mauvaise

Une citation exacte d'un document obsolète, biaisé ou erroné ne rend pas la conclusion fiable. Il faut évaluer la source elle-même — ce que la traçabilité ne fait pas.

Une preuve peut être mal interprétée

Un passage réel peut être sorti de son contexte, soumis à une exception, ou insuffisant pour soutenir la conclusion qu'on lui fait porter.

Le silence n'est pas une preuve d'absence

Si Claude ne trouve pas une information, cela peut signifier qu'elle n'est pas dans le contexte — pas qu'elle n'existe pas. Les deux se traitent différemment.

La validation dépend du risque

Un brouillon marketing et une décision contractuelle n'appellent pas le même niveau de contrôle. Le coût d'une erreur guide la profondeur de la revue.

Calibrer l'effort de contrôle

ENJEU	NIVEAU MINIMAL EXIGÉ	CONTRÔLE
Brouillon interne, exploration	Plausible (3)	Relecture rapide ; les incertitudes doivent être visibles.
Document partagé en interne	Étayé (4)	Source et passage pour chaque point important.
Livrable client, engagement	Validé (5)	Source primaire, contrôle humain, personne compétente identifiée.

Positionnement : Claude et l'IA générative en général

Le risque de contenu faux présenté avec assurance concerne l'IA générative en général, pas seulement Claude. Ce qui est spécifique ici, ce sont les techniques de réduction des hallucinations documentées par Anthropic et, côté API, la fonctionnalité de citations ancrées dans les documents. Un meilleur environnement ou un meilleur modèle améliore la fiabilité ; il ne transforme jamais une génération en preuve. ^{[1][5]}

Avant d'envoyer — six vérifications

- ☐ Chaque chiffre a une source et une date.
- ☐ Les inférences sont marquées comme telles.
- ☐ Les informations manquantes sont listées, pas comblées.
- ☐ Les contradictions entre sources sont signalées.
- ☐ Le niveau exigé par l'enjeu est atteint.
- ☐ Les points à conséquence sont validés par qui de droit.

À RETENIR

La vérification ne rend pas une réponse vraie. Elle rend visible ce qui, dans cette réponse, reste à vérifier par quelqu'un.

GUIDE 04 · 07

Sources principales

Sources Anthropic et NIST consultées le 28 août 2026.

[1] Reduce hallucinations

Claude Platform Docs — permission d'ignorer, citations d'abord, vérification avec rétractation

platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/reduce-hallucinations

[2] AI Risk Management Framework: Generative AI Profile (NIST AI 600-1)

NIST, juillet 2024 — définition de la confabulation

nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[3] Findings from a Pilot Anthropic–OpenAI Alignment Evaluation Exercise

Anthropic Alignment Science, 27 août 2025 — observations de complaisance

alignment.anthropic.com/2025/openai-findings/

[4] Prompting best practices

Claude Platform Docs

platform.claude.com/docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices

[5] Citations

Claude Platform Docs — fonctionnalité API

platform.claude.com/docs/en/build-with-claude/citations

Périmètre de vérification. La fonctionnalité Citations décrite côté Claude Platform / API ne doit pas être confondue avec une garantie générale de vérité du modèle : elle garantit la validité du pointeur, pas celle du contenu pointé. L'étude de complaisance citée en [3] est un exercice pilote croisé entre deux laboratoires, portant sur les modèles disponibles à sa date ; elle documente une tendance observée, non une propriété permanente d'un modèle donné. Les noms de modèles, interfaces et fonctionnalités peuvent évoluer.

VERSION WEB PERMANENTE

consulting.amplify-belgium.com/ressources/une-reponse-convaincante-n-est-pas-une-preuve

DANS LA MÊME COLLECTION

Guide 03 — Travailler en plusieurs passes