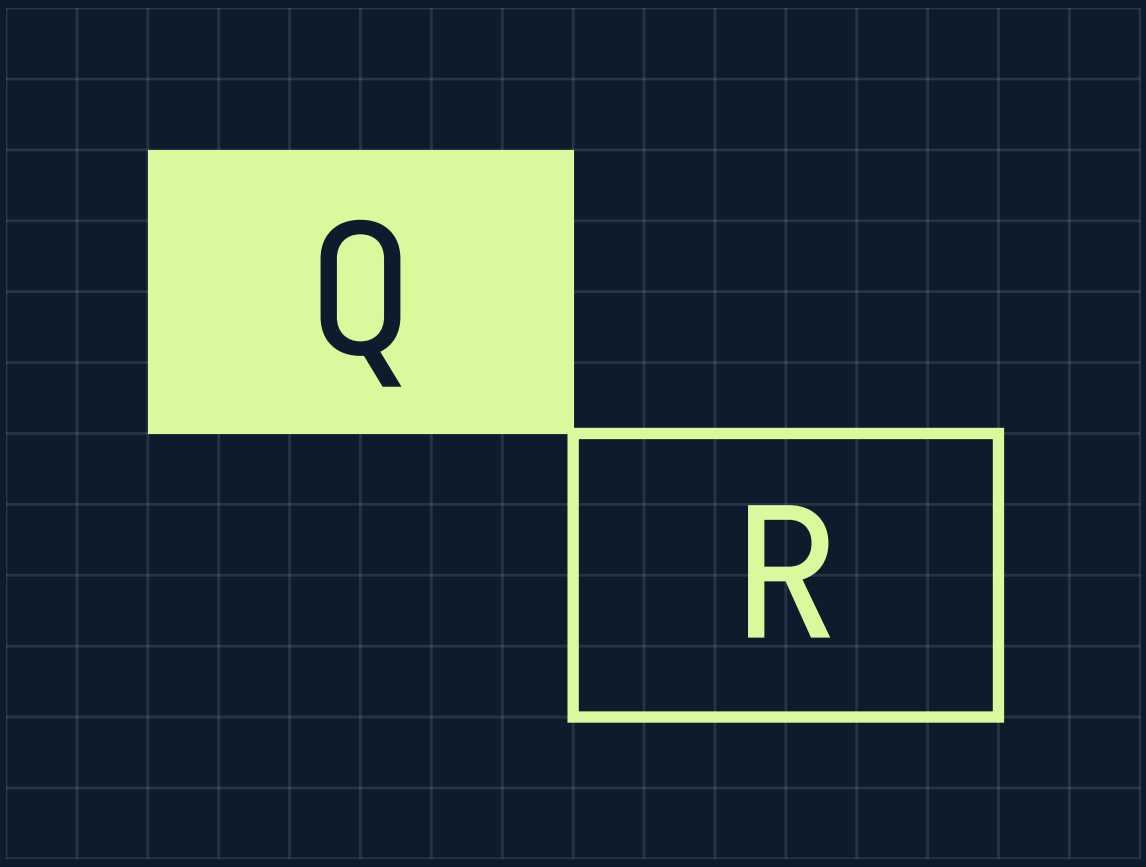


Faire répondre une IA avec vos propres documents

Comment un assistant retrouve puis rédige à partir de vos documents, ce qu'il faut trier avant, et comment juger un essai sur vingt questions dont vous connaissez la réponse.

SOMMAIRE

Ce que contient ce guide

1 « Demandez à vos documents »

2 Comment ça marche, en trois temps

3 Parfois, pas besoin de recherche

4 Trier avant d'interroger

5 Pourquoi une réponse est fausse

6 L'essai sur vingt questions

7 Où les documents sont lus

8 Faire vivre les documents après l'essai

9 Avant et après : un exemple complet

10 Votre premier pas cette semaine

Pour qui. Le dirigeant d'une PME dont les réponses sont dans des dossiers partagés (procédures, contrats, fiches produits) et qui veut savoir ce qu'il faut préparer avant d'essayer un assistant documentaire.

L'idée. L'IA ne « connaît » pas vos documents : elle retrouve des passages puis rédige à partir d'eux ; la qualité dépend d'abord du tri des documents et d'un essai sur des questions dont on connaît la réponse.

Édité par NLQ, avec le soutien d'ALMA. Mis à jour le 19/11/2026. Exemples fictifs ; ce que ce guide dit des outils vient de leur documentation, lue à la date indiquée.

1 « Demandez à vos documents »

La question utile n'est pas « quel outil ? » mais « quels documents, et comment saurai-je que la réponse est juste ? »

Une entreprise de trente personnes (exemple fictif). Ses réponses sont partout : les procédures dans un dossier partagé, les conditions générales dans un autre, les fiches produits dans un tableur, les réponses aux questions fréquentes dans la tête de deux personnes. Chaque nouvel arrivant pose les mêmes questions. Un jour, on présente au dirigeant un assistant « qui répond à partir de tous vos documents ». La démonstration est bluffante.

Trois semaines plus tard, l'assistant cite une procédure remplacée depuis un an. Personne ne s'en aperçoit tout de suite, parce que la réponse était bien écrite.

L'outil n'était pas en cause : il a fait exactement ce qu'il fait, répondre à partir de ce qu'il trouve. La question utile n'était pas « quel outil ? », mais « quels documents lui donne-t-on, et comment saura-t-on que la réponse est juste ? ».

Ce livre blanc explique, en mots simples, comment un assistant répond à partir de vos documents, ce qu'il faut préparer avant, et comment juger un essai avant de décider.

Exercice. Notez les cinq questions qu'on vous pose le plus souvent dans l'entreprise, et pour chacune le document où se trouve la réponse. Si vous ne savez pas lequel, c'est déjà une information.

> **À RETENIR**
> Avant « quel outil ? », demandez « quels documents, et comment saurai-je que la réponse est juste ? ».

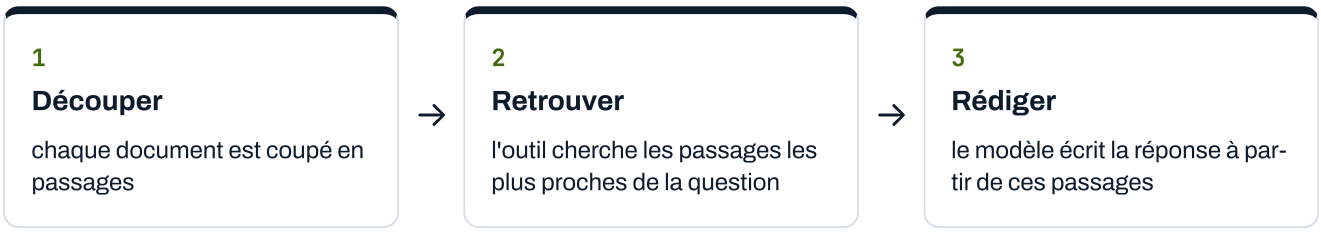
---

2 Comment ça marche, en trois temps

On découpe, on retrouve les passages proches de la question, puis le modèle rédige à partir d'eux

Un assistant ne « connaît » pas vos documents comme un collègue qui les aurait lus. Dans la plupart des outils, il les consulte au moment de la question, selon une méthode qu'on appelle en anglais « retrieval-augmented generation », abrégée RAG : génération augmentée par la recherche. La documentation de Mistral la décrit en deux étapes : retrouver les informations pertinentes dans une base de connaissances, puis les insérer dans la demande faite au modèle, pour qu'il rédige une réponse appuyée sur ce contexte.

Avant ces deux étapes, il y a une préparation : les documents sont découpés en morceaux plus petits, pour que la recherche puisse ensuite isoler les passages les plus proches de la question.



Ce mécanisme explique la plupart des bonnes et des mauvaises surprises. Anthropic, dans un article consacré à cette étape de recherche, note que les méthodes classiques perdent du contexte au moment du découpage, ce qui fait souvent échouer la recherche. Son exemple : un passage qui parle d'une hausse de chiffre d'affaires, mais qui, isolé de son document, ne dit plus de quelle entreprise ni de quelle période il s'agit.

Traduit pour une PME : un paragraphe « Le délai est de 10 jours » n'aide personne s'il a perdu son titre (« Retour des pièces détachées, clients professionnels, conditions 2025 »). D'où l'importance de documents bien titrés, datés, et dont chaque partie dit de quoi elle parle.

> **« L'assistant ne connaît pas vos documents ; il les consulte au moment de la question. »**

> À RETENIR
> Découper, retrouver, rédiger : la réponse ne vaut que ce que valent les passages retrouvés.

3 Parfois, pas besoin de recherche

Quand le corpus est court, on peut le donner en entier à l'assistant

Le même article d'Anthropic rappelle que la solution la plus simple est parfois la meilleure : selon l'éditeur, si une base de connaissances fait moins de 200 000 « jetons », les unités de texte que traite le modèle, soit environ 500 pages, on peut la donner en entier au modèle avec la question, sans mécanisme de recherche. Cette indication concerne ses propres modèles ; chaque outil a ses limites, à vérifier dans sa documentation.

Pour beaucoup de petites entreprises, les documents vraiment de référence (procédures en vigueur, conditions générales, fiches produits actives) tiennent dans quelques dizaines de pages. Des assistants grand public permettent de joindre des fichiers à un espace de travail : l'aide de Gemini indique par exemple qu'un Gem peut recevoir des fichiers comme sources, sous « Connaissances ».

VOTRE SITUATION	LA VOIE LA PLUS SIMPLE
Quelques dizaines de pages de référence, qui changent rarement	Donner les documents en entier à un assistant, dans un espace dédié
Des centaines de documents, ou qui changent souvent	Un outil qui découpe, retrouve et cite ses sources
Des documents dont l'accès dépend de la personne	Un outil qui respecte les droits d'accès existants

Exercice. Estimez la taille de votre corpus utile : pas tout le serveur, seulement les documents qui contiennent des réponses à jour.

> **À RETENIR**
> Quand le corpus utile est court, on commence par le donner en entier ; la recherche vient quand il grandit.

4 Trier avant d'interroger

Un assistant répète ce qu'il trouve, y compris la version périmée

Un assistant répète ce qu'il trouve. S'il trouve trois versions d'une procédure, il peut citer la mauvaise, avec la même assurance que la bonne. Le tri est donc le travail le plus utile, et il ne demande aucune technique.

Pour chaque document candidat, quatre questions :

- **Est-il en vigueur ?** Les versions remplacées vont dans une archive que l'assistant ne lit pas.
- **Qui en est responsable ?** Un document sans responsable ne sera jamais mis à jour.
- **Son titre dit-il de quoi il parle ?** « Procédure_v3_final2 » ne dit rien ; « Retours clients — procédure en vigueur depuis le 01/03/2026 » dit tout.
- **Contient-il des informations que tout le monde ne doit pas voir ?** Salaires, dossiers du personnel, contrats nominatifs : à exclure, ou à réserver à un outil qui respecte les droits d'accès.

Les documents scannés méritent une attention particulière : une image de texte n'est pas toujours lisible par l'outil. Vérifiez qu'on peut y sélectionner le texte, sinon prévoyez une version texte.

Exercice. Prenez vingt documents du dossier le plus consulté. Pour chacun, répondez aux quatre questions dans un tableau. Comptez ceux qui passent les quatre : c'est votre corpus de départ.

> **« Un assistant répète ce qu'il trouve, y compris la version périmée. »**

À RETENIR

Un document entre dans le corpus s'il est en vigueur, a un responsable, un titre clair, et peut être lu par tous ceux qui poseront des questions.

---

5 Pourquoi une réponse est fausse

Cinq causes, chacune avec son symptôme et son remède

Quand une réponse est fausse, la cause est presque toujours l'une de ces cinq. Les reconnaître évite de changer d'outil pour un problème de documents.

CE QUE VOUS VOYEZ	LA CAUSE PROBABLE	LE REMÈDE
La réponse cite une règle ancienne	Une version périmée est dans le corpus	Archiver hors du corpus ; dater les documents
« Je ne trouve pas » alors que l'information existe	Le passage a perdu son contexte, ou le document n'est pas lisible	Titres explicites, intertitres qui répètent le sujet, version texte des scans
Une réponse vague qui mélange deux sujets	La question est ambiguë, ou deux documents se contredisent	Reformuler ; trancher la contradiction dans les documents
Une réponse précise mais inventée	Aucun passage pertinent, et le modèle a comblé le vide	Demander à l'outil de citer ses sources et de dire quand il ne sait pas ; vérifier les citations
Une information que la personne ne devait pas voir	Le corpus ne respecte pas les droits d'accès	Retirer ces documents, ou choisir un outil qui applique les droits existants

La quatrième ligne est la plus dangereuse, parce que la réponse a l'air juste. Un bon réglage consiste à exiger que chaque réponse indique le document et le passage d'où elle vient, et à apprendre à l'équipe à cliquer sur la source avant de faire confiance.

À RETENIR

Une réponse fausse vient le plus souvent des documents ou de la question, pas de l'outil.

6 L'essai sur vingt questions

On juge un assistant sur des questions dont on connaît la réponse, avec un seuil écrit avant

On ne juge pas un assistant sur une démonstration, mais sur vos questions, dont vous connaissez déjà la réponse. Vingt suffisent pour décider de la suite.

Composez-les ainsi :

NOMBRE	SORTE DE QUESTION	CE QU'ON VÉRIFIE
10	Questions courantes, posées chaque semaine	La réponse est juste et cite le bon document
5	Questions rares, dont la réponse est enfouie	L'outil retrouve un passage peu visible
3	Questions dont la réponse n'est dans aucun document	L'outil dit qu'il ne sait pas, au lieu d'inventer
2	Questions où deux documents se contredisent	L'outil signale la contradiction

Pour chaque réponse, une personne qui connaît le métier note l'une de ces cinq mentions : juste et sourcée ; juste sans source ; incomplète ; fausse ; « je ne sais pas » à bon escient.

Avant de lancer l'essai, écrivez votre seuil de décision. Par exemple : on continue si aucune réponse n'est fausse sur les questions courantes et si les trois questions sans réponse donnent bien « je ne sais pas » ; on revoit le tri des documents si plus de deux réponses citent une version ancienne. Un seuil fixé après coup sert surtout à se convaincre.

Exercice. Écrivez vos vingt questions, leurs réponses attendues et le document où chacune se trouve. Cette liste est utile même si vous n'allez jamais plus loin : elle montre où sont vraiment vos réponses.

> **À RETENIR**
> Vingt questions dont vous connaissez la réponse, dont trois sans réponse, et un seuil écrit avant l'essai.

7 Où les documents sont lus

Trois façons de faire tourner l'assistant, et quatre questions à poser pour chacune

Une fois les documents triés et l'essai préparé, reste à choisir où l'assistant les lit. Trois grandes façons de faire existent, chacune avec ses questions.

Un assistant grand public ou professionnel, avec des fichiers joints. Simple à essayer, adapté à un petit corpus. C'est le cas d'un Gem auquel on ajoute des fichiers, par exemple.

Une plateforme qui se connecte à vos espaces de stockage. Certains éditeurs proposent des fonctions gérées qui chargent, découpent et cherchent dans vos documents à votre place, ou des connecteurs vers des espaces existants : la documentation de Mistral mentionne ainsi des « Libraries » et des connecteurs vers Google Drive ou SharePoint.

Un outil installé sur vos propres serveurs. Certains logiciels s'installent chez vous et interrogent les documents sans les envoyer chez un éditeur. La documentation d'Open WebUI, par exemple, le présente comme une plateforme d'IA auto-hébergée, conçue pour fonctionner entièrement hors ligne, avec une fonction de recherche dans les documents ; il a sa propre licence, à lire avant tout usage. Ces outils demandent des compétences pour l'installation, la mise à jour et la sécurité.

Pour chacune, quatre questions à poser, par écrit :

- Où sont copiés mes documents, et combien de temps ?
- Servent-ils à autre chose qu'à répondre à mes questions ?
- Une personne peut-elle, par l'assistant, lire un document qu'elle ne peut pas ouvrir dans le dossier ?
- Chaque réponse cite-t-elle son document et son passage ?

La troisième question est souvent oubliée. Un assistant qui lit tout le serveur avec un seul compte peut rendre visible à tous ce qui ne l'était qu'à quelques-uns.

> **À RETENIR**
> Choisissez l'endroit après le tri et l'essai, pas avant ; et vérifiez que l'assistant respecte les droits d'accès.

8 Faire vivre les documents après l'essai

L'assistant reflète vos documents et vieillit avec eux ; trois habitudes le gardent juste

Un assistant documentaire ne se règle pas une fois pour toutes. Il reflète vos documents : le jour où une procédure change sans que le document soit mis à jour, il répond avec l'ancienne règle. Trois habitudes suffisent à le garder juste.

Garder la liste des questions sans réponse. Chaque fois que l'assistant répond « je ne sais pas », ou qu'une personne signale une réponse fausse, la question est notée avec la date. Cette liste vaut de l'or : elle montre ce qui manque dans vos documents. Une fois par mois, le responsable du corpus la parcourt et décide, pour chaque ligne, s'il faut écrire un document, en corriger un, ou accepter que la question reste hors du périmètre.

Dater les révisions. Chaque document du corpus porte une date de dernière vérification et le nom de son responsable. Une procédure qui n'a pas été relue depuis un an n'est pas forcément fausse, mais elle mérite un coup d'œil avant d'être citée à un nouvel arrivant.

Rejouer les vingt questions. Après chaque changement important, un document ajouté, une procédure réécrite, un réglage modifié, on rejoue l'essai sur les vingt questions. Une réponse qui était juste et qui devient fausse se voit tout de suite, au lieu d'être découverte par un client.

QUAND	QUI	QUOI
À chaque réponse fausse signalée	La personne qui l'a vue	Note la question, la réponse et la date
Chaque mois	Le responsable du corpus	Parcourt la liste et décide document par document
À chaque changement de procédure	Le responsable du document	Met à jour le document et sa date
Après un changement important	Le responsable du corpus	Rejoue les vingt questions

Exercice. Désignez dès maintenant le responsable du corpus. Sans ce nom, les trois habitudes ne tiendront pas plus de quelques semaines.

> **« Un assistant documentaire reflète vos documents : il vieillit avec eux. »**

> **À RETENIR**
> Une liste des questions sans réponse, des documents datés, et les vingt questions rejouées après chaque changement.

9 Avant et après : un exemple complet

Le même projet, préparé avec le tri et l'essai, change de taille

Reprenons l'entreprise du début (exemple fictif).

AVANT	APRÈS
« Un assistant qui répond à partir de tous vos documents »	**« Un assistant qui répond aux questions des nouveaux arrivants sur les retours, les délais et les garanties »**
Tout le serveur, trois versions par procédure	**34 documents en vigueur, titrés et datés, chacun avec un responsable**
Démonstration sur des questions choisies par le vendeur	**Essai sur vingt questions de l'entreprise, seuil écrit avant**
Réponses sans source	**Chaque réponse cite son document ; l'équipe apprend à cliquer**
Accès de l'outil à tout	**Aucun document réservé dans le corpus**

Le projet a rétréci, et il est devenu vérifiable. Si l'essai est concluant, on élargira le corpus document par document, avec la même règle de tri.

> **À RETENIR**
> Un projet bien préparé commence petit, et grandit document par document.

10 Votre premier pas cette semaine

Vingt questions, leurs réponses, les documents où elles se trouvent

- [ ] J'ai noté les cinq questions qu'on me pose le plus, et le document de chaque réponse.
- [ ] J'ai trié vingt documents avec les quatre questions.
- [ ] J'ai estimé la taille de mon corpus utile.
- [ ] J'ai écrit mes vingt questions d'essai, avec leurs réponses attendues.
- [ ] J'ai écrit mon seuil de décision avant tout essai.

NOTES

Sources et portée

Mistral — RAG quickstart — consulté le 19/11/2026.

Anthropic — Contextual Retrieval — consulté le 19/11/2026. Les chiffres de performance de cet article ne sont pas repris ici.

Aide Gemini — Utiliser des Gems — consulté le 19/11/2026.

Open WebUI — documentation — consulté le 19/11/2026, cité comme exemple d'outil auto-hébergé, sans recommandation.

Ce livre blanc résume des documentations d'éditeurs ; il ne compare ni ne note aucun produit. Le tri, la grille d'essai et l'exemple (fictif) sont des propositions. Les limites et fonctions des outils changent : vérifiez-les dans leur documentation avant de choisir.

VOTRE PROCHAINE ÉTAPE

Un échange gratuit

Apportez vos vingt questions et la liste de vos documents : on regarde ensemble ce qu'un essai pourrait montrer.

Réserver un échange gratuit

www.nl-q.com/rendez-vous/

1

Vous choisissez un créneau sur la page de rendez-vous du site.

2

Vous recevez le lien de visioconférence créé pour l'échange.

3

On part de votre situation. Venez avec la liste, même partielle : les questions sans document connu sont souvent les plus instructives.

Édité par NLQ — NL-Q (www.nl-q.com), avec le soutien d'ALMA. Édition du 19/11/2026. Exemples fictifs ; aucune promesse de résultat ; ce qui est dit des textes et des outils cités vient des sources datées dans les notes.