Newsletter N° 18 – Avril 2026

RAG : comment l'IA peut travailler

Avec vos propres documents

La technique qui transforme un chatbot généraliste en expert de votre domaine.

Vous testez Claude ou ChatGPT depuis des mois, vous êtes impressionné par leur culture générale, mais une frustration revient toujours : ils ne connaissent pas vos documents. Vos jurisprudences utiles, vos modèles d'actes, vos notes de réunion, vos procédures, vos archives clients. Toute cette matière qui fait votre quotidien professionnel reste inaccessible au modèle. Vous pouvez lui copier-coller des extraits, mais cela ne change rien : à la prochaine conversation, il aura tout oublié.

Cette limite n'est pas une fatalité. La technique pour la dépasser existe, elle est mature, et elle commence à s'industrialiser dans toutes les professions exigeantes. Elle porte un nom un peu rébarbatif : RAG, pour Retrieval-Augmented Generation.

Newsletter N° 18 – Avril 2026

Génération augmentée par récupération. Derrière cette traduction maladroite se cache l'une des avancées les plus utiles de l'IA appliquée au monde professionnel.

Le problème que le RAG résout

Pour comprendre l'intérêt du RAG, il faut d'abord bien voir ce qu'il vient corriger.

Un grand modèle de langage a deux limites structurelles quand il s'agit de votre matière professionnelle.

Première limite : il ne connaît que ce qu'il a appris pendant son entraînement. Si vos documents internes n'étaient pas dans ses données d'entraînement (et ils ne l'étaient pas), il ne sait rien sur eux.

Deuxième limite : sa mémoire de conversation est volatile. Ce que vous lui dites dans une session disparaît à la session suivante. Aucune accumulation, aucun apprentissage continu.

Conséquence pratique : pour le faire travailler sur votre matière, 2 options s'offrent à vous, toutes 2 insatisfaisantes. Soit vous lui collez à chaque fois des extraits dans le prompt, ce qui est fastidieux, limité par la longueur du

contexte, et impossible à industrialiser. Soit vous le faites entraîner sur vos données (fine-tuning), ce qui est coûteux, lent, et fige le modèle à un instant donné.

Le RAG offre une troisième voie, beaucoup plus élégante : on garde le modèle tel quel, on construit à côté une base de connaissances qui contient vos documents, et on installe un système de pont entre les deux.

Quand vous posez une question, le système va d'abord chercher dans votre base les passages pertinents, puis les fournit au modèle pour qu'il formule sa réponse en s'appuyant dessus.

Le principe expliqué simplement

Le fonctionnement d'un système RAG peut se résumer en 4 étapes.

Première étape, en amont : la mise en base. On prend tous les documents que l'on veut rendre interrogeables (PDF, fichiers Word, mails, notes, jurisprudences) et on les découpe en passages de taille raisonnable, généralement quelques paragraphes. Chaque passage est ensuite transformé en vecteur, c'est-à-dire en une représentation numérique de son sens. Ces vecteurs sont stockés dans une base spécialisée, appelée base vectorielle.

Deuxième étape, à chaque requête : la recherche. Quand vous posez une question, votre question est elle aussi transformée en vecteur. Le système cherche alors dans la base les passages dont les vecteurs sont les plus proches de celui de votre question. Cette proximité numérique correspond, en pratique, à une proximité de sens. Les passages les plus pertinents remontent.

Troisième étape : l'injection. Les passages retrouvés sont insérés dans le prompt qui sera envoyé au modèle, avec votre question. Le modèle reçoit donc à la fois la question et la matière première pour y répondre.

Quatrième étape : la génération. Le modèle formule sa réponse en s'appuyant sur les passages fournis. La consigne qui lui est donnée précise généralement : ne réponds qu'à partir des passages ci-dessous, et cite les sources que tu utilises.

Ce que ça change concrètement

L'avantage principal du RAG, c'est qu'il transforme la relation entre l'IA et votre matière. L'IA cesse d'être une encyclopédie généraliste qui ne connaît pas votre métier, pour devenir un assistant qui s'appuie sur vos propres ressources.

Pour un cabinet d'avocats, cela peut vouloir dire un système qui interroge l'ensemble des notes internes du cabinet, des consultations passées, des

conclusions déjà rédigées sur des sujets similaires. Vous demandez : ai-je déjà eu un dossier proche de celui-ci ?, et le système ressort les trois précédents les plus pertinents avec les passages clés.

Pour une étude notariale, cela peut être un système qui croise les modèles d'actes maison, les notes doctrinales archivées, les correspondances avec les administrations. La question quelle clause utilise-t-on habituellement dans ce cas trouve une réponse instantanée.

Pour un service juridique d'entreprise, le RAG peut donner accès à l'historique des contrats, aux notes de la direction, aux décisions du comité juridique. L'effet sur la mémoire institutionnelle est immédiat : ce qui se perdait dans les boîtes mail individuelles devient interrogeable.

Pour un professionnel libéral isolé, c'est l'accès à sa propre matière accumulée. Tous les dossiers traités, les recherches déjà faites, les modèles construits au fil des ans, deviennent une base vivante.

Newsletter N° 18 – Avril 2026

Les vrais avantages, par rapport au fine-tuning

Une question revient souvent : pourquoi ne pas plutôt entraîner le modèle directement sur ses documents (fine-tuning) ? La réponse tient en plusieurs points.

Coût et temps : un fine-tuning sérieux coûte cher et prend du temps.

Un système RAG peut être mis en place pour une fraction du coût et fonctionne dès la première heure.

Fraîcheur : quand vous fine-tunez un modèle, il connaît vos données à la date du fine-tuning. Tout ce que vous ajoutez ensuite lui est inconnu jusqu'au fine-tuning suivant. Avec le RAG, il suffit d'ajouter le document dans la base : il est immédiatement interrogeable.

Traçabilité : un modèle fine-tuné vous donne une réponse, sans toujours pouvoir indiquer d'où vient l'information. Un système RAG bien conçu cite ses sources : tel paragraphe vient de tel document. C'est crucial pour un usage professionnel.

Contrôle : avec le RAG, vous pouvez à tout moment retirer un document de la base, et l'IA n'y aura plus accès. Avec un modèle fine-tuné, l'information est cuite dans les paramètres, impossible à retirer proprement.

Sécurité : vos documents ne quittent jamais votre base vectorielle.

Selon l'architecture choisie, ils peuvent même rester intégralement sur votre infrastructure. Le modèle externe ne voit que les passages spécifiquement remontés pour répondre à une question donnée.

Les limites à connaître

Le RAG n'est pas une solution magique. Plusieurs difficultés méritent d'être anticipées.

La première, c'est la qualité du retrieval. Si le système ne retrouve pas les bons passages, la réponse sera médiocre, voire fausse. Or, le retrieval n'est pas parfait. Il dépend de la qualité du découpage des documents, de la pertinence de la transformation en vecteurs, et de la formulation de la question. Une question mal posée peut ramener des passages hors sujet.

La deuxième, c'est la hiérarchie des sources. Un système RAG basique traite toutes les sources sur un pied d'égalité. Or, dans votre base, un texte de loi n'a pas le même poids qu'une note interne. Une consultation ancienne dépassée n'a pas le même poids qu'une consultation récente. Il faut construire cette hiérarchie.

La troisième, c'est la qualité initiale des documents.

Un RAG ne fait pas de miracles : si vos archives sont mal organisées, mal nommées, mal indexées, le système en héritera. Il accélère ce qui était déjà accessible, il ne rattrape pas le désordre de fond.

La quatrième, c'est le coût du maintien. Une base RAG vit.

Il faut y ajouter les nouveaux documents, retirer les périmés, vérifier que le découpage reste adapté. Sans cette maintenance, le système se dégrade.

Par où commencer

Pour qui souhaite explorer concrètement, plusieurs voies sont possibles.

La voie la plus accessible : les outils intégrés. Plusieurs offres grand public proposent désormais une fonctionnalité projet ou workspace permettant d'attacher des documents à une conversation. Claude permet d'attacher des PDF à un projet, ChatGPT propose des GPT personnalisés avec base documentaire. Ce n'est pas un RAG industriel, mais c'est une porte d'entrée.

La voie intermédiaire : les plateformes spécialisées. Plusieurs solutions, certaines françaises, permettent de construire un RAG sans coder. On dépose

les documents, on configure quelques paramètres, on obtient une interface de questions-réponses. Tarification généralement à l'usage.

La voie sur mesure : le développement spécifique. Pour des besoins métier précis, faire développer une solution dédiée par un prestataire reste l'option la plus puissante. Coût plus élevé, mais maîtrise totale de l'architecture et des garanties.

Quelle que soit la voie choisie, le bon point de départ est toujours le même : identifier un cas d'usage précis et limité, le tester, en mesurer la valeur réelle, puis élargir. Vouloir tout RAGifier d'un coup est la meilleure façon de ne rien obtenir.

Une transformation silencieuse

Le RAG est probablement l'une des avancées les plus structurantes de l'IA appliquée au monde professionnel, et pourtant l'une des moins médiatisées. Pas de démos spectaculaires, pas de buzz sur les réseaux. Juste, peu à peu, une intégration discrète dans les outils métier.

Pour les professions qui vivent de leur matière documentaire (droit, comptabilité, conseil, médecine, recherche), c'est un changement de paradigme.

Newsletter N° 18 – Avril 2026

La connaissance accumulée cesse d'être une ressource passive, archivée, peu interrogeable. Elle devient une ressource vivante, mobilisable à la demande.

Ceux qui prendront le temps de structurer leur matière maintenant prendront une avance considérable dans les deux prochaines années. Ceux qui attendront que ce soit prêt à l'emploi devront rattraper un retard difficile à combler.

Passez une excellente journée

Gabriel PAPP

gpappAI.com