SPOTLAB / ANALYSE P15

Qualité des connaissances et RAG : pourquoi une bonne IA peut donner une mauvaise réponse

Diagnostiquer la fiabilité d'un RAG de la source jusqu'à la réponse et gouverner les connaissances dans la durée.

Analyse / P15

Novembre 2026

Sources consultées au 2 novembre 2026

ANALYSE P15

Sommaire

01	Trois enseignements pour les décideurs	3
02	Le vrai problème : une chaîne de connaissances, pas seulement un modèle	4
03	Six couches de qualité à diagnostiquer	5
04	Cas illustratif : deux politiques de crédit, une mauvaise filiale	6
05	Passer d'une correction intuitive à une amélioration démontrée	8
06	Gouverner la connaissance dans la durée	10
07	Les cinq décisions à prendre	11
08	Limites et méthode	12
09	Références	12

FICHE DE L'ANALYSE

ÉDITEUR
Spotwork.AI — Spotlab, Research & Insights

RÉFÉRENCE
P15

TYPE
Analyse documentaire

COLLECTION
Données & connaissances

SOURCES CONSULTÉES
2 novembre 2026

PUBLICATIONS LIÉES
I0-02 · I0-03 · P14

EN BREF

Quand un assistant d'entreprise cite la mauvaise procédure ou applique une règle périmée, changer de modèle n'est pas nécessairement la bonne solution. Le défaut peut venir du document, de son extraction, de l'index, des droits d'accès, de la recherche ou du contexte effectivement transmis au modèle. Fiabiliser le RAG suppose de diagnostiquer le défaut à son origine — et de démontrer que la correction tient dans le temps.

> *La qualité des connaissances, leur actualité et leur provenance déterminent en grande partie la fiabilité des réponses.*
>
> CE QU'IL FAUT RETENIR

01 Trois enseignements pour les décideurs

1

Une source retrouvée n'est pas forcément une preuve utilisable.

Elle doit être pertinente pour l'entité, la version, la période et la question ; une citation cliquable ne garantit ni son exactitude ni son autorité.

2

La qualité se perd à chaque transformation.

Extraction de PDF, segmentation, métadonnées, entités, relations et résumés peuvent tous altérer ou masquer un fait. Les architectures GraphRAG ajoutent des possibilités de liaison, mais aussi des objets extraits à vérifier [3].

3

Une réponse fiable est aussi capable de ne pas conclure.

La recherche de Google sur le « sufficient context » distingue les cas où les preuves permettent de répondre des cas insuffisamment étayés ; l'abstention doit être mesurée, sans devenir systématique [1].

Synthèse exécutive

PROBLÈME

La fiabilité d'une application RAG dépend du modèle, mais aussi de la chaîne documentaire qui alimente la réponse. L'ingestion d'un PDF peut perdre une négation ; un index peut manquer une pièce ; un filtre peut inclure la mauvaise filiale ; une réponse peut citer un texte dépassé.

THÈSE

Séparer la qualité intrinsèque des sources, l'intégrité de leurs transformations, la qualité de leur récupération, la suffisance du contexte et l'exactitude de la réponse. La traçabilité relie les étapes ; elle ne garantit pas la vérité du contenu.

CONTRIBUTION

Un cadre de diagnostic à six couches et un contrat de tests reproductible. Une synthèse et une proposition de méthode, non une étude expérimentale ni des résultats observés chez un client.

La qualité ne se résume ni à une mesure unique du RAG, ni à un score de confiance calculé par le modèle. **La sécurité des autorisations s'impose à chaque étape de la chaîne.**

02 Le vrai problème : une chaîne de connaissances, pas seulement un modèle

Le RAG (retrieval-augmented generation) récupère des informations extérieures au modèle pour éclairer une réponse. Cette approche peut améliorer la pertinence et la vérifiabilité, mais elle ne transforme pas automatiquement chaque document récupéré en fait incontestable.

Une procédure peut être authentique mais périmée ; une note peut concerner la mauvaise filiale ; une phrase peut avoir perdu sa négation pendant l'extraction ; un passage décisif peut être absent du contexte. Des travaux sur les conflits entre sources et sur les connaissances obsolètes montrent que ces situations demandent des traitements distincts [4] [5].

> TROIS QUESTIONS À POSER SÉPARÉMENT
>
> Ainsi, trois questions doivent être posées séparément : avons-nous la bonne connaissance ? l'avons-nous correctement retrouvée ? l'avons-nous correctement utilisée ?

Ce que les travaux démontrent — et ce qu'ils ne démontrent pas

A Évaluer à plusieurs niveaux

La revue publiée en 2026 dans *SN Computer Science* par Knollmeyer et al. synthétise douze études et distingue cinq dimensions d'évaluation — pertinence du contexte, fidélité, pertinence de la réponse, exactitude et qualité des citations —, ainsi que les types de questions et les domaines d'application [2]. Elle constitue un cadre conceptuel ; elle ne démontre pas qu'une configuration unique surpasse toutes les autres.

B Suffisance du contexte et abstention

Les travaux de Google Research présentés à ICLR 2025 [1] observent que des modèles propriétaires répondent bien lorsque le contexte est suffisant, mais produisent souvent une réponse erronée plutôt que de s'abstenir lorsqu'il ne l'est pas ; des modèles ouverts hallucinent ou s'abstiennent fréquemment, même avec un contexte suffisant. La capacité d'abstention dépend donc de la suffisance de l'information et du comportement du modèle. Ces résultats restent liés aux modèles et corpus évalués, non transposables automatiquement à tout système d'information d'entreprise.

C Sources contradictoires et obsolètes

Les travaux de Google Research sur les contradictions entre sources [4] proposent une typologie des conflits et montrent qu'ils appellent des comportements différents ; le benchmark HoH, publié à l'ACL 2025 [5], montre que des informations obsolètes dégradent l'exactitude, même lorsque l'information à jour est aussi récupérée. Ils ne fixent pas de seuil universel de fraîcheur : une procédure RH peut nécessiter une mise à jour quasi immédiate ; un document historique peut rester valide pendant des années.

02 / CE QUE LES TRAVAUX DÉMONTRENT — SUITE

D

Représentations structurées

Microsoft documente les objets créés par le pipeline d'indexation GraphRAG — Document, TextUnit, Entity, Relationship, Covariate, Community et Community Report [3]. Ces représentations offrent des possibilités de recherche, mais imposent des vérifications de fidélité et de provenance des objets extraits. Il serait trompeur de présenter un graphe comme vrai simplement parce qu'il a été extrait de documents.

03 Six couches de qualité à diagnostiquer

Le cadre ci-dessous est une proposition méthodologique Spotwork.AI de diagnostic, et non une norme publiée ni la preuve d'un composant déjà industrialisé.



Couche	Défaut typique	Contrôle attendu	Preuve minimale
1. Source	Procédure dépassée, document sans autorité	Propriétaire, statut, date de validité et source de référence	Identifiant source, propriétaire, statut, date, droits
2. Transformation	Tableau mal extrait, négation perdue	Comparaison avec l'original, journal d'extraction et d'indexation	Extrait original, version du parseur, différentiel
3. Représentation	Entité ou relation erronée dans le graphe	Validation d'entités, de liens et de provenance	Relation sourcée, passage d'origine, version
4. Récupération	Passage pertinent absent du top-k	Rappel sur jeu annoté, filtres temporels et droits d'accès	Contexte de référence, top-k, requête, ACL
5. Contexte	Éléments incomplets ou contradictoires	Test de suffisance, priorité des sources et détection des conflits	Besoins couverts, lacunes et conflits explicités
6. Réponse	Conclusion non étayée ou mauvaise attribution	Exactitude métier, justification par affirmation, abstention contrôlée	Vérification métier, citations et décision

Cette séparation permet d'éviter une correction coûteuse au mauvais endroit. Si le passage exact existe mais n'est pas récupéré, changer de LLM risque de masquer le défaut sans le résoudre. Si le passage est présent mais mal interprété, améliorer le moteur de recherche ne suffit pas.

> **UNE ERREUR PEUT RELEVER DE PLUSIEURS COUCHES**
>
> La taxonomie vise le diagnostic causal. Une source périmée correctement récupérée et fidèlement synthétisée reste une réponse métier fausse. Un score de « faithfulness » élevé peut donc être trompeur si les sources elles-mêmes sont invalides.

04 Cas illustratif : deux politiques de crédit, une mauvaise filiale

Un responsable risque interroge un assistant sur le seuil d'exception applicable à une demande de crédit. L'outil récupère une politique de la filiale A, datée de 2024, alors qu'une version de 2026 s'applique à la filiale B. L'assistant produit une réponse fluide avec citation, mais la référence documentaire est inadaptée au dossier.

Le diagnostic n'est pas « hallucination du modèle » par défaut. Il faut vérifier : quel périmètre la question visait-elle ? quelles métadonnées portaient les documents ? quel filtre a été appliqué ? quelle version a réellement été fournie au modèle ? La réponse correcte peut nécessiter de demander la filiale, d'exclure les versions abrogées, ou de signaler que la preuve disponible est insuffisante.

SCÉNARIO FICTIF Cet exemple est fictif et n'illustre aucun déploiement client réel ni résultat mesuré.

Approfondissement : le cas de contrôle

Un contrôleur compare une limite de risque pour une filiale et une date données. Trois sources coexistent : un document historique de 2024 relatif à la filiale A ; un document applicable à la filiale B depuis mars 2026 ; une note d'interprétation non validée. Le moteur récupère en premier une clause de la filiale A et formule une réponse catégorique.

Ce qu'un système maîtrisé teste séparément	Couche	Question de contrôle
Existence de la bonne source	Source	La politique de la filiale B est-elle présente, valide et faisant autorité ?
Validité des métadonnées, extraction du montant et de l'entité	Transformation	La filiale, la date d'effet et le seuil ont-ils été correctement extraits ?
Filtres d'accès et bonne section	Récupération	Le filtre d'entité et de période a-t-il été appliqué ; l'utilisateur y a-t-il droit ?
Présence de la clause dans le contexte	Contexte	La clause applicable figure-t-elle dans ce que le modèle a reçu ?
Raisonnement final	Réponse	Le modèle a-t-il hiérarchisé les sources et signalé l'ambiguïté ?

04 / INTERPRÉTER L'ÉCHEC ET ÉPROUVER CHAQUE CONTRÔLE

En cas d'ambiguïté non résolue, une question de clarification ou une abstention expliquée est préférable à une conclusion artificielle.

CONTRE-EXEMPLE VOLONTAIRE

Après mise à jour de l'index, la nouvelle source est récupérée, mais le modèle continue de retenir l'ancien montant figurant dans une autre pièce du contexte. Ce n'est plus une erreur de récupération seule ; c'est un problème de hiérarchisation et d'utilisation du contexte. **Il faut un jeu de tests multicouche et ne pas s'arrêter au Recall@k.**

Séparer l'erreur de périmètre de l'erreur de validité

Le cas initial cumule deux défauts : la politique récupérée concerne la mauvaise filiale *et* une version ancienne. Pour savoir quel contrôle a manqué, il faut éprouver chaque défaut isolément. Les deux variantes ci-dessous sont hypothétiques.

Variante	Pièce récupérée à tort	Ce qui doit l'écarter	Nature de l'erreur
Cas initial	Politique 2024 de la filiale A	Filtre d'entité et filtre de validité	Périmètre et validité temporelle cumulés
Variante 1 — bonne filiale, version remplacée	Politique 2024 de la filiale B, remplacée en mars 2026	Métadonnées de date d'effet et de statut « en vigueur »	Validité temporelle
Variante 2 — version en vigueur, mauvaise filiale	Politique 2026 de la filiale A	Métadonnée d'entité et filtre de périmètre lié au dossier	Périmètre

Si la variante 1 échoue seule, la correction porte sur la gestion des versions ; si la variante 2 échoue seule, sur les métadonnées d'entité et les filtres. Un système qui ne réussit que le cas initial par hasard, par exemple parce que le bon document est aussi le plus récent, n'a pas démontré la maîtrise de ces deux contrôles.

Trois défauts, trois contrôles distincts

PÉRIMÈTRE
Métadonnée d'entité et filtre lié au dossier

VALIDITÉ
Date d'effet, statut « en vigueur » et versions remplacées

HIÉRARCHISATION
Règle de priorité des sources dans le contexte transmis

05 Passer d'une correction intuitive à une amélioration démontrée

Avant de modifier une configuration, une équipe devrait conserver un jeu de questions représentant ses usages : questions simples, multi-documents, versions contradictoires, cas sans réponse, accès restreints et données récemment modifiées. Un protocole utile mesure chaque étape et la réponse finale.

L'étude de synthèse publiée en 2026 dans *SN Computer Science* souligne que la qualité d'un RAG doit être évaluée à plusieurs niveaux : pertinence du contexte, fidélité aux sources, exactitude, pertinence de la réponse et qualité des citations. Un score global ne localise pas nécessairement le défaut [2].

La démarche recommandée consiste à :

01 Décrire

Décrire l'incident avec la question, la réponse, le contexte reçu et les références aux documents.

02 Attribuer

Attribuer l'erreur à une ou plusieurs couches, sans accuser automatiquement le modèle.

03 Corriger

Corriger de manière ciblée la source, l'extraction, l'index, le filtre, le contexte ou la génération.

04 Rejouer

Rejouer un jeu de non-régression et vérifier également les coûts, les délais, les droits et le comportement d'abstention.

05 Surveiller

Surveiller les changements du corpus et les permissions pour ne pas réintroduire une erreur corrigée.

> *L'objectif n'est pas de répondre davantage, à tout prix : c'est d'augmenter les réponses justes et étayées, sans faire progresser les réponses dangereusement assurées.*

Protocole de test reproductible et borné

Le contrat de jeu de tests

Constituer les cas avec identifiant stable, question, entité, période de validité, droits d'accès, documents autorisés, passages de référence, réponse métier attendue ou type d'abstention. Séparer un jeu de développement d'un jeu de décision. Inclure des questions simples, multi-pièces, temporalisées, contradictoires, sans réponse, interdites, ainsi qu'un événement de remplacement de document.

05 / PROTOCOLE DE TEST — SUITE

Mesures proposées par niveau

Niveau	Mesures proposées
Source	Part de documents avec propriétaire et validité vérifiables ; suivi du retrait des versions abrogées
Récupération	Couverture des passages pertinents à k fixé (Recall@k si annotations disponibles), qualité des filtres d'accès et de la validité temporelle ; examen des erreurs hors top-k
Contexte	Part des cas où l'ensemble récupéré contient les preuves nécessaires, conflits détectés, justesse de la hiérarchie des sources
Réponse	Exactitude métier sur les cas annotés, fidélité des affirmations à leur source, citations vérifiables, abstentions justifiées et sur-refus
Exploitation	Délai complet, coût par question, durée d'actualisation du corpus, incidents de droits, stabilité après réindexation

Les métriques automatisées ne sont pas un oracle

Les travaux RAGAS [9] et la revue de 2026 [2] étayent la séparation des mesures de récupération et de génération. Les juges IA doivent toutefois être calibrés contre une référence humaine ; leurs versions, consignes et divergences doivent être tracées. Un test de récupération devrait aussi vérifier que la question exige réellement le corpus : une prépublication de mai 2026 (Liu et al., Purdue University et Microsoft Research) montre que de nombreux benchmarks RAG contiennent des questions auxquelles un modèle peut déjà répondre de mémoire [11].

Stabilité et limites

Un cas est une question ou condition de test distincte ; une répétition est une exécution du même cas, dans une configuration déterminée. Définir avant exécution le nombre de répétitions selon le risque et la complexité, un plafond de tests, de coût et de durée, et les conditions d'arrêt sur incident critique. Ne pas extrapoler une absence d'erreurs observées à une garantie de risque nul ; publier les intervalles d'incertitude lorsque des conclusions statistiques sont formulées.

Décision de mise en service

GO LIMITÉ

La qualité métier et les contrôles d'accès satisfont les seuils définis à l'avance.

CORRECTION PUIS RETEST

Le problème est localisable et sans incident critique.

NO-GO

Accès possible à un document interdit, références critiques non vérifiables ou erreurs métier graves au-delà du seuil.

Les seuils sont à déterminer avec le métier, et non présentés comme des standards universels.

06 Gouverner la connaissance dans la durée

La qualité documentaire ne se limite pas à une note calculée une fois. Un document doit avoir un responsable, une version, un périmètre d'application, des droits d'accès, des dates de validité et une politique de mise à jour ou retrait. Quand la source évolue, les extraits, résumés, embeddings et relations qui en dépendent doivent être réévalués selon les impacts réels.

Le W3C fournit avec PROV-O un vocabulaire pour décrire des entités, activités et agents de provenance ; cela facilite la traçabilité, sans établir que le contenu est vrai. ISO/IEC 25012 décrit pour sa part un modèle général de qualité des données structurées : une référence utile, pas une certification de qualité d'un système RAG [6] [7].

L'accès doit également être vérifié au moment de la récupération et de la restitution. Un index contenant des informations confidentielles, même exactes, ne doit pas les divulguer à un utilisateur non autorisé [8].

La fiche minimale d'une connaissance de référence

IDENTIFIANT CANONIQUE · PROPRIÉTAIRE · ORIGINE · DATE DE PUBLICATION · DATE D'EFFET · DATE DE RETRAIT · STATUT · JURIDICTION · ENTITÉ · PÉRIMÈTRE · SENSIBILITÉ · RÈGLE D'ACCÈS · EMPREINTE OU VERSION · DÉPENDANCES D'INDEXATION

Certains champs peuvent être non pertinents selon le cas ; ils doivent être explicitement évalués, pas inventés.

Une politique de rafraîchissement par classe de sources

Le rafraîchissement ne se confond pas avec un nouveau crawl général à chaque question.

Mode	Usage type
Événementiel	Quand une version change
Périodique	Données dynamiques
À la demande	Dossier exceptionnel
Stable	Références historiques

Un retrait doit propager les invalidations de fragments, embeddings, résumés, caches et relations dérivées lorsqu'ils sont concernés. L'absence de mise à jour est une décision qui doit rester contrôlable.

Sécurité et intégrité

L'OWASP attire l'attention sur l'empoisonnement documentaire, le contrôle des droits sur les fragments récupérés, l'isolation et la surveillance [8]. La provenance et le hachage participent à l'intégrité ; ils ne remplacent ni la validation de la source ni les droits. Le profil IA générative du NIST [10] fournit un cadre volontaire de gestion des risques et d'évaluation, sans certifier une solution particulière.

07 Les cinq décisions à prendre

Décision	Question à trancher
1. Cadrer la connaissance	Quelles sources font autorité pour chaque usage ?
2. Qualifier et versionner	Quelles règles garantissent la validité métier et temporelle ?
3. Diagnostiquer	À quelle couche se situe chaque erreur significative ?
4. Évaluer	Quels tests démontrent la pertinence, l'exactitude et la sûreté ?
5. Entretenir	Qui actualise les sources et déclenche les revalidations ?

L'enjeu stratégique est de transformer un stock de documents en **connaissance professionnelle gouvernée**, et non de multiplier les documents dans un index.

Mise en œuvre progressive

ÉTAPE A
Qualifier 1 à 2 corpus
Définir les sources faisant autorité, le responsable, les droits, les dates et la règle de retrait.

ÉTAPE B
Baseline et diagnostic
Constituer les cas métier, rejouer la configuration actuelle et attribuer les erreurs par couche.

ÉTAPE C
Améliorer de façon ciblée
Corriger indexation, représentation, filtres et contexte sans tout réécrire.

ÉTAPE D
Contrôle continu
Automatiser les tests critiques lors des changements documentaires, surveiller les écarts et organiser l'escalade humaine.

Le résultat attendu est une décision documentée et réversible, pas une affirmation de fiabilité absolue. L'approche est compatible avec des architectures RAG classiques, hybrides ou GraphRAG ; il reste nécessaire de vérifier les bénéfices empiriques dans le contexte métier et réglementaire visé.

08 Limites et méthode

Cette publication est une synthèse documentaire assortie d'une proposition de méthode. Aucun test propriétaire Spotwork.AI n'a été conduit pour cet article. Les sources citées ne prouvent ni l'efficacité d'une implémentation spécifique ni l'existence de fonctionnalités livrées. Un score de fidélité à des sources erronées peut être élevé ; une bonne citation ne remplace donc pas la vérification de l'autorité et de l'exactitude des documents. Les juges IA peuvent aussi varier et nécessitent une calibration humaine.

Les exemples et seuils sont fictifs ou à définir. Aucun audit de conformité, test de sécurité, preuve de performance ou certification de produit n'est revendiqué. Les études citées ont leurs propres corpus et protocoles. Les recommandations de gouvernance ne constituent ni un avis juridique ni une adaptation certifiée de standards normatifs.

09 Références

[1] Joren et al., Google Research — Sufficient Context: A New Lens on Retrieval Augmented Generation Systems (ICLR 2025).
research.google/pubs/sufficient-context-a-new-lens-on-retrieval-augmented-generation-systems/

[2] Knollmeyer et al. — Evaluating Retrieval Augmented Generation: A Comprehensive Review, *SN Computer Science* 7, 576 (2026).
link.springer.com/article/10.1007/s42979-026-05134-x

[3] Microsoft — GraphRAG, Indexing Dataflow et modèle de connaissances.
microsoft.github.io/graphrag/index/default_dataflow/

[4] Cattan et al., Google Research (2025) — Detecting and Addressing Conflicting Sources in Retrieval-Augmented LLMs.
research.google/pubs/dragged-into-a-conflict-detecting-and-addressing-conflicting-sources-in-retrieval-augmented-llms/

[5] Ouyang et al. — HoH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on RAG (ACL 2025).
aclanthology.org/2025.acl-long.301/

[6] W3C — PROV-O: The PROV Ontology, Recommandation (30 avril 2013).
w3.org/TR/prov-o/

[7] ISO/IEC 25012:2008 — Data quality model (confirmée en 2025).
iso.org/standard/35736.html

[8] OWASP Cheat Sheet Series — RAG Security Cheat Sheet.
cheatsheetseries.owasp.org/cheatsheets/RAG_Security_Cheat_Sheet.html

[9] Es et al. — Ragas: Automated Evaluation of Retrieval Augmented Generation (arXiv, 2023).
arxiv.org/abs/2309.15217

[10] NIST AI 600-1 — AI RMF: Generative Artificial Intelligence Profile (juillet 2024).
doi.org/10.6028/NIST.AI.600-1

[11] Liu et al. — Generating Leakage-Free Benchmarks for Robust RAG Evaluation (prépublication arXiv, mai 2026).
arxiv.org/abs/2605.08838

Nature des preuves : articles de recherche et prépublications, documentation fournisseur, recommandation W3C, norme ISO/IEC, guide OWASP et cadre volontaire du NIST. Sources consultées au 2 novembre 2026.

Spotwork.AI

L'IA professionnelle de confiance

Mieux travailler aujourd'hui,
construire votre intelligence de demain.

SPOTLAB · RESEARCH & INSIGHTS · ANALYSE P15

Qualité des connaissances et RAG : pourquoi une bonne IA peut donner une mauvaise réponse

Diagnostiquer de la source jusqu'à la réponse, gouverner dans la durée

© 2026 SPOTWORK.AI