Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Fabrique sémiotique hybride et évaluation stratifiée des systèmes RAG sur SciFact

Christophe Denis[1,2]
(1) UMMISCO, Sorbonne Université, France
(2) IHPST, Université Paris 1 Panthéon-Sorbonne, France
christophe.denis@sorbonne-universite.fr

RÉSUMÉ

Cet article propose, dans le cadre de la fabrique sémiotique hybride, une méthodologie stratifiée et traçable d'évaluation des systèmes RAG sur SciFact. *Hermes-HSF* articule des métriques classiques de recherche d'information avec des indicateurs opérationnels portant sur l'ancrage, les citations, les réponses de repli et la stabilité. L'analyse montre qu'une meilleure récupération documentaire ne garantit pas une meilleure réponse, ce qui justifie une évaluation des médiations internes de la chaîne RAG.

ABSTRACT

Hybrid Semiotic Fabric and Stratified Evaluation of RAG Systems on SciFact

This paper proposes a traceable and stratified methodology for evaluating RAG systems on SciFact, grounded in the hybrid semiotic fabric. *Hermes-HSF* combines standard information-retrieval metrics with operational indicators for grounding, citation visibility, fallback answers, and stability. The results show that better retrieval does not necessarily yield better answers, supporting an evaluation of the internal mediations of the RAG chain.

MOTS-CLÉS : RAG, évaluation, récupération documentaire, reclassement, SciFact, LLM.

KEYWORDS: RAG, evaluation, retrieval, reranking, SciFact, LLM.

1 Problématique et hypothèse de travail

L'objectif de l'atelier *EvalLLM 2026* est d'interroger l'évaluation des grands modèles de langage (LLM) dans des dispositifs complexes de traitement documentaire et informationnel, en particulier dans les systèmes de génération augmentée par récupération (RAG). Cette question se pose dans un contexte où les protocoles d'évaluation demeurent encore largement construits à partir de corpus anglophones. L'enjeu est double :

— Les systèmes RAG constituent des architectures hybrides, articulant recherche documentaire, reconstruction contextuelle et génération. Leur évaluation ne peut donc se réduire à une mesure unique portant sur la seule réponse finale. Une telle évaluation suppose de distinguer la récupération documentaire, la génération et le fonctionnement d'ensemble (Yu *et al.*, 2024; Lewis *et al.*, 2020).
— Les cadres d'évaluation disponibles restent largement élaborés à partir de corpus d'évaluation anglophones, ce qui limite leur capacité à rendre compte de la diversité linguistique et

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 99–110.

contextuelle des usages effectifs et à apprécier correctement les performances en contexte multilingue (Thakur *et al.*, 2025).

Les travaux consacrés à l'évaluation des systèmes RAG regroupent fréquemment les métriques en deux grandes catégories : les métriques de recherche documentaire, d'une part, et les métriques de contexte et de réponse, d'autre part. Cette distinction est nécessaire, mais elle demeure insuffisante dès lors que l'évaluation est rapportée à la seule sortie terminale du système. Une réponse finale, même jugée correcte, ne permet pas à elle seule de comprendre le processus qui la produit, ni les opérations de sélection des sources, de construction du contexte et de formulation de la réponse qui en conditionnent la forme et l'intelligibilité.

La contribution de cet article consiste à proposer une méthodologie stratifiée et traçable d'évaluation des systèmes RAG. Cette méthodologie ne réduit pas l'analyse à la correction finale de la réponse produite. Elle articule des métriques classiques de recherche d'information avec des indicateurs opérationnels propres au protocole *Hermes-HSF*, afin de rendre comparables plusieurs niveaux de la chaîne RAG : récupération documentaire, construction du contexte, génération, attribution aux sources, réponses de repli et stabilité entre exécutions.

L'enjeu n'est donc pas d'ajouter une métrique isolée, mais de rendre observable la manière dont les différentes médiations du pipeline contribuent à la qualité ou à la dégradation de la réponse finale. En ce sens, *Hermes-HSF* est moins conçu comme un évaluateur unique que comme un dispositif d'expérimentation permettant de comparer plusieurs configurations internes d'un système RAG.

C'est cette limite, à savoir l'impossibilité de rapporter l'évaluation à la seule sortie finale, qui conduit nos travaux à un déplacement théorique vers la philosophie du langage. Dès lors qu'il ne s'agit plus de réduire l'évaluation à la seule conformité apparente d'une sortie, il devient nécessaire d'interroger les conditions dans lesquelles cette sortie prend sens, devient interprétable et peut être tenue pour pertinente dans un cadre d'usage déterminé.

Une telle inflexion trouve un appui dans l'œuvre de Wittgenstein. Comme il l'écrit dans les *Recherches philosophiques*, au §43, « la signification d'un mot est son emploi dans le langage » (Wittgenstein, 2014). Dans le cadre de cet article, cette référence n'a pas pour fonction d'introduire une théorie générale du langage, mais de soutenir un déplacement méthodologique précis : dans un système RAG, une réponse ne peut être évaluée indépendamment des usages documentaires, des contraintes de contexte, des procédures de récupération et des formes d'attribution qui rendent cette réponse interprétable.

L'hypothèse structurante de nos travaux est donc que l'évaluation des systèmes fondés sur des LLM, et plus encore des pipelines RAG, doit porter sur la chaîne de médiations qui rend une réponse possible, intelligible et discutable. C'est dans cette perspective que nous mobilisons le concept de fabrique sémiotique hybride (Denis, 2025). Celui-ci vise à penser un régime de production du sens dans lequel humains, corpus, dispositifs techniques, procédures de sélection et contraintes d'énonciation interviennent conjointement. Une sortie produite par un système de type LLM ou RAG ne peut donc être réduite à un résultat terminal évalué pour lui-même : elle doit être rapportée à l'ensemble des médiations qui en conditionnent la possibilité, la forme et l'intelligibilité.

La plateforme informatique *Hermes-HSF* a été conçue pour donner à cette hypothèse un cadre opératoire et expérimental. Elle vise à rendre observables, traçables et comparables les médiations documentaires, contextuelles et énonciatives à l'œuvre dans les interactions avec les LLM. Elle a pour fonction d'instrumenter les transformations successives par lesquelles une requête, un corpus, une opération de récupération documentaire, une contrainte de prompt ou une procédure de restitution

reconfigurent la réponse produite.

La présente contribution se concentre sur le module de *Hermes-HSF* consacré à l'évaluation comparative de pipelines RAG. Il ne s'agit donc pas seulement de comparer plusieurs configurations techniques, mais de mettre à l'épreuve cette hypothèse structurante en distinguant explicitement plusieurs niveaux d'analyse : l'accès aux sources, la structuration du contexte transmis au générateur, le régime de réponse produit, la visibilité des appuis documentaires et le degré d'ancrage effectif de la sortie dans les matériaux mobilisés.

Pour préciser cette méthodologie stratifiée, nous distinguons quatre dimensions d'évaluation des systèmes RAG, qui ne doivent pas être confondues avec un score unique portant sur la seule réponse finale :

1. **La récupération documentaire.** Elle évalue si les documents ou passages récupérés sont pertinents pour répondre à la requête. Cette dimension peut être mesurée par des métriques classiques de recherche d'information, telles que la précision, le rappel, le rang moyen réciproque ou la précision moyenne normalisée, selon la structure des annotations disponibles.
2. **La qualité de la réponse produite.** Elle concerne la correction factuelle, la complétude, la concision ou l'adéquation à une réponse de référence. Dans un système RAG, cette qualité ne peut toutefois pas être interprétée indépendamment des sources mobilisées : une réponse apparemment correcte peut être produite à partir d'un contexte documentaire incomplet ou mal exploité.
3. **L'ancrage de la réponse dans les sources.** Cette dimension, souvent désignée par les notions de *faithfulness* ou de *groundedness*, vise à déterminer si les affirmations générées sont effectivement soutenues par les passages récupérés, ou si elles introduisent des éléments non justifiés par le contexte fourni au modèle.
4. **L'attribution aux sources.** Elle évalue la capacité du système à relier les affirmations de la réponse aux documents ou passages qui les justifient. Cette dimension est décisive dans les usages documentaires, car elle conditionne la vérifiabilité de la réponse et la possibilité, pour un lecteur humain, de contrôler la chaîne probatoire.

La plateforme *Hermes-HSF* s'inscrit dans cette pluralité de dimensions. Sa contribution ne consiste pas à les remplacer par un score unique, mais à les articuler dans une méthodologie stratifiée et traçable. Le protocole permet ainsi d'observer séparément la récupération documentaire, la construction du contexte, la génération, l'attribution aux sources, les réponses de repli et la stabilité entre exécutions. L'objectif est de rendre diagnostiquable le comportement du pipeline : une dégradation peut provenir du rappel documentaire, de la sélection des passages, de la formulation du prompt, de la génération elle-même ou encore de l'absence d'attribution fiable.

La section suivante présente les objectifs et l'architecture générale de *Hermes-HSF*. Le protocole expérimental est ensuite exposé, avant la présentation des résultats, leur discussion, les limites de l'étude et la conclusion.

2 Objectif et architecture de la plateforme *Hermes-HSF*

La plateforme *Hermes-HSF* a été conçue pour donner à l'hypothèse de la fabrique sémiotique hybride un cadre opératoire et expérimental. Son objectif n'est pas seulement d'obtenir une sortie du modèle,

mais de rendre observables, traçables et comparables les transformations qui relient un état du dispositif, une forme de textualisation, une production symbolique et un effet interprétable. Elle vise ainsi à étudier, dans un cadre contrôlé, les médiations par lesquelles une production issue d'un grand modèle de langage devient intelligible, actionnable ou discutable.

L'architecture générale de *Hermes-HSF* repose sur une même hypothèse : un modèle ne doit pas être observé comme une instance isolée produisant un résultat final, mais comme un composant inséré dans une chaîne de médiations contrôlées. La Figure 1 en propose une vue synthétique. Elle met en évidence un noyau commun de médiations, à partir duquel se déploient deux modules expérimentaux complémentaires : un module de couplage perception–action en environnement discret et un module d'évaluation comparative de pipelines de génération augmentée par récupération.

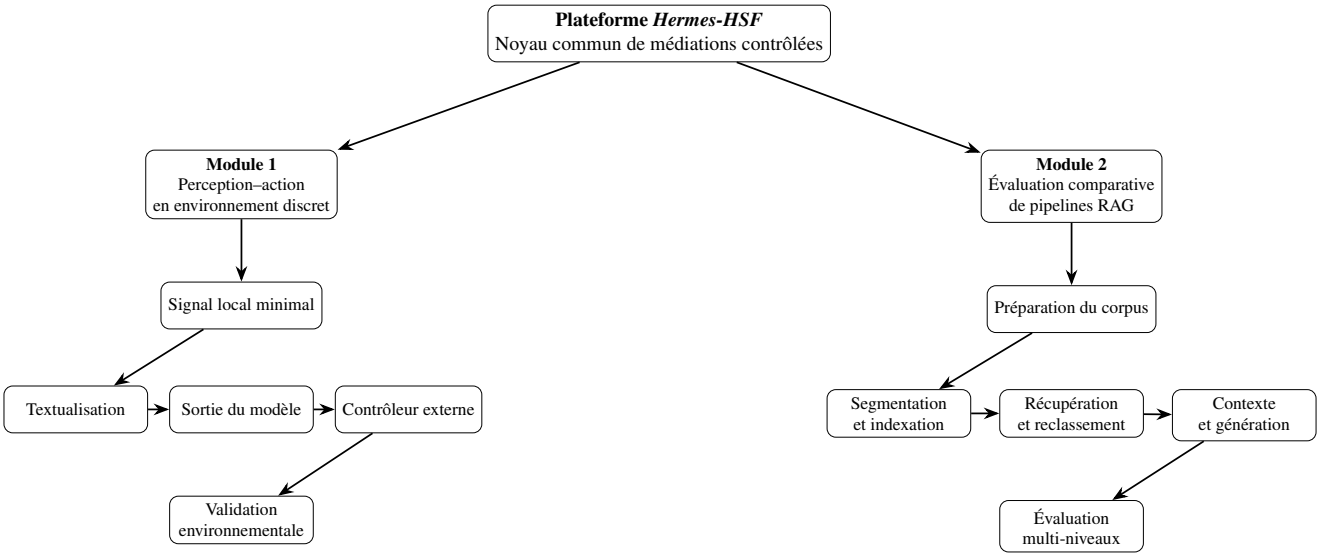


FIGURE 1 – Architecture générale de la plateforme *Hermes-HSF*. La plateforme articule un noyau commun de médiations contrôlées et deux modules expérimentaux complémentaires : un module perception–action en environnement discret et un module d'évaluation comparative de pipelines RAG.

La plateforme articule ainsi deux modules expérimentaux complémentaires :

— **Un module de couplage minimal entre perception et action dans un environnement discret.** Dans ce premier module, le modèle n'a jamais accès à un état global du monde. Il reçoit un signal local minimal, textualisé sous une forme strictement contrôlée, puis produit un signe d'action dans un espace de réponses borné. Cette sortie n'est pas exécutée directement : elle est interprétée par un contrôleur externe, qui la confronte aux contraintes de l'environnement avant d'enregistrer son effet. Ce module permet ainsi d'étudier, dans un cadre fortement contrôlé, la chaîne reliant signal, textualisation, décision symbolique et validation environnementale.

— **Un module d'évaluation comparative de pipelines de génération augmentée par récupération.** Dans ce second module, objet de la présente contribution, l'environnement n'est plus principalement un espace discret d'action minimale, mais un corpus documentaire préparé selon plusieurs représentations. La plateforme articule alors préparation des documents, segmentation en unités textuelles, indexation, récupération documentaire, reclassement éventuel, reconstruction du contexte, génération et évaluation. Le dispositif permet de comparer

plusieurs configurations de pipeline en distinguant explicitement l'accès aux sources, la structuration du contexte transmis au générateur, le régime de réponse produit, la visibilité des appuis documentaires et le degré d'ancrage effectif de la sortie dans les matériaux mobilisés.

Ces deux modules relèvent d'une même logique architecturale. Dans le premier cas, il s'agit de suivre les médiations qui relient perception locale, décision symbolique et effet environnemental. Dans le second, il s'agit de suivre les médiations qui relient corpus, sélection documentaire, reconstruction contextuelle et production de réponse. Dans les deux cas, *Hermes-HSF* vise à déplacer l'analyse depuis la seule performance terminale vers les opérations intermédiaires qui conditionnent la production du sens.

La section suivante présente le protocole expérimental retenu pour l'étude de ce second module.

3 Protocole expérimental

Le protocole expérimental retenu vise à mettre à l'épreuve le second module de la plateforme *Hermes-HSF*, consacré à l'évaluation comparative de pipelines de génération augmentée par récupération. Il ne s'agit pas seulement d'en mesurer la performance globale, mais de distinguer plusieurs niveaux de fonctionnement du dispositif : l'accès aux sources, la structuration du contexte transmis au générateur, la forme des réponses produites et la stabilité des sorties d'une exécution à l'autre.

La Figure 2 en donne une vue d'ensemble. Le protocole articule la préparation du corpus, la constitution de deux campagnes expérimentales, une double représentation documentaire, plusieurs configurations de pipeline et, enfin, une évaluation multi-niveaux des sorties produites.

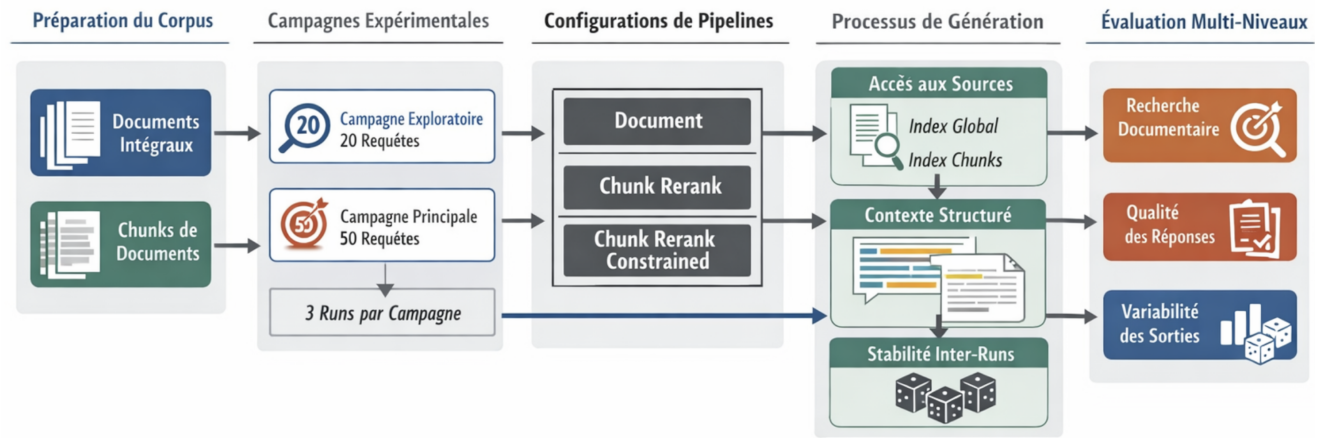


FIGURE 2 – Vue d'ensemble du protocole expérimental mis en œuvre pour le module RAG de *Hermes-HSF*.

3.1 Données mobilisées

Les données mobilisées dans cette contribution sont issues de SciFact, utilisé ici à travers BEIR (*Benchmarking IR*), un cadre unifié d'évaluation pour la recherche d'information (Thakur *et al.*, 2021). SciFact est un jeu de données de vérification d'assertions scientifiques composé d'environ 1 400 énoncés rédigés par des experts, associés à des résumés d'articles contenant des éléments de preuve

annotés. Dans notre cas, son intérêt est de fournir un corpus documenté, un ensemble de requêtes stabilisées et un cadre reproductible pour comparer plusieurs configurations de pipeline.

Ce choix ne signifie pas que SciFact constituerait un horizon suffisant de l'évaluation. Construit en anglais, il ne saurait épuiser la diversité linguistique et contextuelle des usages effectifs. Nous l'utilisons donc comme un environnement de test contrôlé, destiné à mettre à l'épreuve notre hypothèse structurante sur un corpus standardisé.

3.2 Préparation du corpus

Les documents issus de SciFact sont importés depuis BEIR, puis convertis dans un format local compatible avec l'environnement expérimental de *Hermes-HSF*. Les requêtes et les jugements de pertinence associés sont conservés de manière à permettre une comparaison directe entre les sorties du pipeline et les documents attendus.

Le corpus est ensuite préparé selon une double représentation. D'un côté, les documents sont conservés comme unités intégrales. De l'autre, ils sont segmentés en unités textuelles de taille contrôlée avec recouvrement partiel. Cette double représentation permet de comparer deux régimes de récupération documentaire : un régime fondé sur le document entier et un régime fondé sur des unités textuelles plus fines. Dans la configuration retenue avec Ollama, l'index documentaire repose sur une représentation TF–IDF, tandis que l'index des unités textuelles repose sur des embeddings (Karpukhin *et al.*, 2020). Le calcul d'embeddings a ainsi été réservé au niveau de l'unité textuelle, afin d'éviter les difficultés liées à l'indexation vectorielle de documents entiers trop longs.

3.3 Campagnes et configurations

Deux campagnes ont été menées. La première repose sur un sous-ensemble de 20 requêtes et joue un rôle exploratoire et qualitatif. La seconde constitue le cœur de l'étude comparative et repose sur 150 requêtes. Dans les deux cas, les expériences sont exécutées à cinq reprises selon des initialisations distinctes, afin d'introduire une mesure explicite de variabilité.

Trois configurations principales sont comparées sur la campagne à 150 requêtes. La première, *document*, opère une récupération au niveau du document entier (Lewis *et al.*, 2020). La seconde, *chunk_rerank*, opère une récupération au niveau des unités textuelles et ajoute une étape de reclassement (Nogueira & Cho, 2019). La troisième, *chunk_rerank_constrained*, conserve la même structure documentaire que la précédente, mais ajoute une contrainte explicite au niveau du prompt de génération, en demandant au modèle de répondre à partir des seules sources fournies et d'en rendre visibles les appuis. La campagne à 20 requêtes est conduite avec cette dernière configuration.

Config.	**Unité**	**Traitement**	**Effet évalué**
`doc.`	Document entier	Contexte large	Accès au document complet
`chunk`	Unités textuelles	Récupération et reclassement	Granularité fine et pertinence locale
`chunk constr.`	Unités textuelles	Reclassement et ancrage contraint	Visibilité des sources et fidélité

TABLE 1 – Configurations RAG comparées dans le protocole expérimental.

`chunk constr.` abrège `chunk_rerank_constrained`.

Ce tableau vise à rendre explicite la logique comparative du protocole. Les configurations ne sont pas seulement distinguées par leurs performances finales, mais par les médiations qu'elles modifient : granularité documentaire, sélection des passages, reclassement, construction du contexte et contrainte de génération. Cette distinction est centrale pour interpréter les résultats, car une amélioration observée à un niveau du pipeline peut ne pas se traduire mécaniquement à un autre niveau.

3.4 Chaîne de traitement

Le pipeline suit une logique en plusieurs temps. Une requête textuelle est d'abord soumise au module de récupération documentaire, qui interroge soit l'index documentaire, soit l'index segmenté selon la configuration étudiée. Dans certaines configurations, la hiérarchie initiale est ensuite retravaillée par une opération de reclassement. Les éléments retenus sont alors convertis en contexte textuel explicite, sérialisés avec leurs identifiants de source et intégrés à un prompt destiné au générateur. Enfin, la réponse produite fait l'objet d'une évaluation multi-niveaux.

Ce protocole permet ainsi de suivre, pour une même requête, la chaîne complète de médiations reliant la formulation initiale, l'accès aux sources, la structuration contextuelle et le régime final d'énonciation.

3.5 Niveaux d'évaluation et métriques

Le protocole d'évaluation retenu dans ce travail vise à distinguer plusieurs dimensions du pipeline RAG, afin de ne pas rabattre son analyse sur un score unique. Les métriques décrites ici correspondent aux principales dimensions mobilisées dans l'analyse des résultats : qualité de la recherche documentaire, ancrage de la réponse dans le contexte, présence explicite de citations, fréquence des réponses de repli, longueur des sorties et stabilité entre exécutions. Certaines d'entre elles, telles que *ans_len*, interviennent principalement dans le commentaire interprétatif du régime d'énonciation plutôt que dans la synthèse tabulaire centrale.

La qualité de la recherche documentaire est évaluée à l'aide de trois indicateurs classiques de la recherche d'information (Manning *et al.*, 2008; Järvelin & Kekäläinen, 2002). Le *Hit@*k mesure la présence d'au moins un document pertinent parmi les k premiers résultats. Le *MRR* (*Mean Reciprocal Rank*) renseigne sur le rang du premier document pertinent et permet d'apprécier la rapidité avec laquelle une information utile apparaît dans le classement. Le *nDCG* (*Normalized Discounted Cumulative Gain*) évalue la qualité globale du classement en pondérant les résultats par leur position.

Afin de caractériser plus finement le régime de réponse, plusieurs indicateurs complémentaires sont mobilisés. L'ancrage de la réponse dans le contexte (*ans_sup*) fournit une approximation du degré d'appui de la réponse sur les passages transmis au générateur. La présence explicite de citations (*cit_pres*) mesure la visibilité des marques de source dans la réponse. L'indicateur de réponse de repli (*fb_flag*) détecte les cas où le système déclare ne pas pouvoir répondre ou adopte une formulation générique de défausse.

Nous mobilisons également la longueur des réponses (*ans_len*), qui renseigne sur les transformations du régime discursif, ainsi que la stabilité entre exécutions (*ans_stab*), qui estime la robustesse pratique

des configurations face à la variabilité générative.

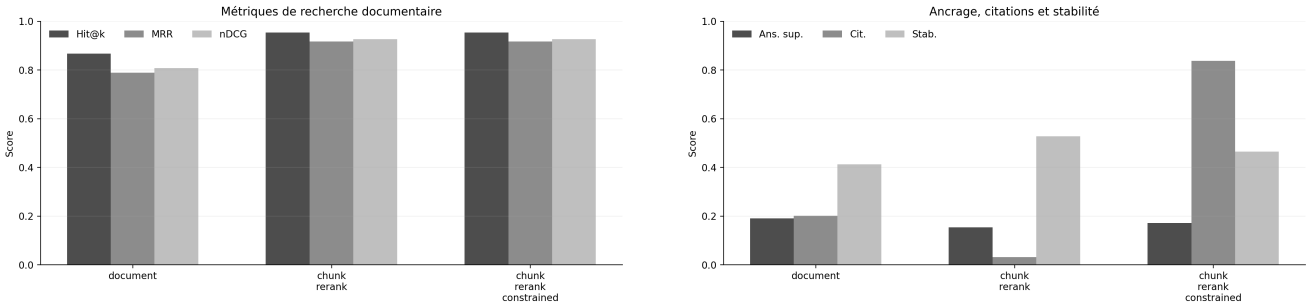
L'ensemble de ces métriques répond à une exigence méthodologique centrale : dissocier ce qui relève de la récupération documentaire, de l'ancrage contextuel, de la visibilité des sources, du régime de réponse et de la robustesse entre exécutions. L'évaluation ne vise donc pas à produire un score synthétique unique, mais à rendre observables plusieurs dimensions de fonctionnement du pipeline, parfois convergentes, parfois divergentes.

4 Résultats expérimentaux

Cette section présente deux campagnes complémentaires. La première, menée sur 20 requêtes, soutient une analyse exploratoire du pipeline contraint. La seconde constitue le cœur de l'étude comparative : elle porte sur 150 requêtes, chaque configuration étant exécutée cinq fois avec des initialisations distinctes. L'objectif n'est pas d'établir une hiérarchie générale entre architectures de RAG, mais de comparer plusieurs configurations selon des dimensions distinctes : qualité de la recherche documentaire, ancrage de la réponse dans le contexte, visibilité des sources, réponses de repli et stabilité entre exécutions. La longueur des réponses est mobilisée comme indicateur complémentaire du régime d'énonciation.

Configuration	n	**Hit@k**	**MRR**	**nDCG**	**Ans. sup.**	**Cit.**	**Fallback**	**Stab.**
20q / *chunk_rerank_constrained*	100	0.650	0.600	0.613	0.168	0.780	0.000	0.442
150q / *document*	750	0.867	0.788	0.808	0.190	0.201	0.009	0.412
150q / *chunk_rerank*	750	0.953	0.917	0.926	0.154	0.031	0.003	0.527
150q / *chunk_rerank_constrained*	750	0.953	0.917	0.926	0.171	0.837	0.032	0.464

TABLE 2 – Résumé des principaux résultats expérimentaux. Hit@k, MRR et nDCG évaluent la qualité de la recherche documentaire. Le tableau reporte également l'ancrage de la réponse dans le contexte, la présence explicite de citations, la fréquence des réponses de repli et la stabilité entre exécutions.



(a) Métriques de recherche documentaire.

(b) Ancrage, citations et stabilité.

FIGURE 3 – Comparaison des pipelines selon, à gauche, les principales métriques de recherche documentaire sur la campagne à 150 requêtes agrégée sur cinq exécutions et, à droite, trois dimensions centrales du régime de réponse : l'ancrage dans le contexte, la présence explicite de citations et la stabilité entre exécutions. Les configurations fondées sur des unités textuelles avec reclassement dominent sur les métriques de recherche documentaire, tandis que la configuration contrainte accroît fortement la visibilité des sources sans rejoindre la configuration documentaire sur l'ancrage dans le contexte. La fréquence des réponses de repli, reportée dans le tableau de synthèse, n'est pas représentée ici.

La Table 2 et la Figure 3 font apparaître une dissociation nette entre les dimensions du pipeline. Sur la campagne à 150 requêtes, *chunk_rerank* et *chunk_rerank_constrained* dominent *document* en recherche documentaire : Hit@k de 0.953, MRR de 0.917 et nDCG de 0.926, contre 0.867, 0.788 et 0.808 pour *document*. Le passage à une récupération fondée sur des unités textuelles avec reclassement améliore donc fortement l'identification et le classement des éléments pertinents.

Cette amélioration ne se prolonge toutefois pas mécaniquement au niveau de la génération. *document* obtient le meilleur score d'ancrage, avec 0.190, contre 0.154 pour *chunk_rerank* et 0.171 pour *chunk_rerank_constrained*. La configuration la plus performante du point de vue de la récupération documentaire n'est donc pas celle qui produit les réponses les plus fortement ancrées dans le contexte transmis au générateur.

La comparaison entre *chunk_rerank* et *chunk_rerank_constrained* précise cette dissociation. Les deux pipelines présentent des performances identiques en recherche documentaire, mais leurs régimes de réponse diffèrent fortement. Lorsque la contrainte de prompt est activée, la présence explicite de citations passe de 0.031 à 0.837 et la longueur moyenne des réponses augmente de 263.04 à 771.02 caractères. En revanche, l'ancrage effectif ne progresse que modestement, de 0.154 à 0.171. Le prompt contraint modifie donc la forme de la réponse et la visibilité de ses appuis documentaires, sans amélioration proportionnelle de son ancrage.

La configuration *document* présente un profil singulier : elle cite peu, recourt légèrement aux réponses de repli et produit des sorties de longueur intermédiaire, mais elle conserve le meilleur score d'ancrage. Elle montre qu'un contexte moins performant du point de vue de la récupération documentaire peut néanmoins conduire à des réponses lexicalement plus proches du matériau mobilisé dans la génération.

La campagne courte à 20 requêtes confirme cette tendance : la configuration contrainte produit des réponses longues et fréquemment citées, sans garantir un fort recouvrement avec les sources injectées.

Les scores de stabilité, compris entre 0.412 et 0.527, indiquent une variabilité réelle mais non dominante entre exécutions. Les écarts observés entre pipelines ne semblent donc pas réductibles à

une initialisation particulière. La configuration *chunk_rerank* apparaît comme la plus stable, devant *chunk_rerank_constrained* et *document*.

Ces résultats confirment l'intérêt d'une évaluation stratifiée : l'amélioration d'un niveau du pipeline ne se traduit pas mécaniquement par une amélioration de l'ensemble de la chaîne RAG. *Hermes-HSF* rend visibles ces dissociations en distinguant récupération, construction du contexte, génération, citations, réponses de repli et stabilité.

Cette dissociation rejoint l'hypothèse de la fabrique sémiotique hybride : la réponse produite résulte d'un agencement entre corpus, indexation, récupération, prompt, génération et attribution. L'évaluation doit donc porter sur cette chaîne de médiations, et non seulement sur la conformité apparente de la réponse à une référence.

Les résultats ne permettent pas d'ordonner les pipelines selon une hiérarchie univoque. Les configurations avec unités textuelles et reclassement dominent si l'on privilégie la recherche documentaire. *chunk_rerank_constrained* l'emporte si l'on privilégie la visibilité explicite des sources. *document* conserve le meilleur score si l'on privilégie l'ancrage de la réponse au contexte. L'intérêt principal de l'étude réside donc dans la mise en évidence de dissociations qui justifient une évaluation stratifiée des systèmes RAG.

5 Discussion

Les résultats obtenus montrent l'intérêt d'une évaluation stratifiée des systèmes RAG. Les configurations fondées sur des unités textuelles avec reclassement améliorent nettement les métriques de recherche documentaire, mais cette amélioration ne se traduit pas mécaniquement par un meilleur ancrage de la réponse dans le contexte. Inversement, la configuration *document*, moins performante du point de vue de la récupération, conserve le meilleur score d'ancrage. Cette dissociation confirme qu'un pipeline RAG ne peut pas être évalué à partir d'un indicateur unique.

La comparaison entre *chunk_rerank* et *chunk_rerank_constrained* précise ce résultat. Les deux configurations obtiennent les mêmes performances en recherche documentaire, mais la contrainte de génération modifie fortement le régime de réponse : elle accroît la visibilité des citations et la longueur des sorties, sans produire une amélioration proportionnelle de l'ancrage effectif. La visibilité des sources ne doit donc pas être confondue avec leur usage réel dans la génération.

Ces résultats soutiennent l'hypothèse méthodologique défendue dans cet article. Une réponse RAG n'est pas le simple produit d'un modèle génératif isolé ; elle résulte d'un agencement entre corpus, segmentation, indexation, récupération, reclassement, construction du contexte, prompt, génération et attribution aux sources. La fabrique sémiotique hybride permet de nommer cette distribution des opérations et d'en faire un objet d'évaluation.

L'apport de *Hermes-HSF* est alors de rendre ces médiations observables et comparables. L'évaluation ne porte plus seulement sur la conformité apparente de la réponse finale, mais sur les transformations qui la rendent possible. C'est ce déplacement qui permet de diagnostiquer des situations où la récupération s'améliore sans que la réponse progresse, où les citations deviennent visibles sans garantir un meilleur ancrage, ou encore où la stabilité varie indépendamment des performances documentaires.

Cette lecture invite à déplacer l'évaluation des systèmes RAG d'une logique de score global vers

une logique d'analyse située des médiations. Elle ne remplace pas les métriques existantes, mais les articule à des indicateurs opérationnels capables de montrer où se produisent les gains, les pertes et les effets ambivalents dans la chaîne de traitement.

6 Limites et considérations éthiques

Cette étude présente trois limites principales. Premièrement, les résultats portent sur *SciFact*, benchmark principalement anglophone, et ne permettent pas de conclure à la robustesse du protocole dans des environnements multilingues ou sur des corpus disciplinaires plus hétérogènes. Deuxièmement, les indicateurs opérationnels produits par *Hermes-HSF* ne constituent pas une mesure exhaustive de la qualité d'une réponse : ils servent à distinguer plusieurs niveaux de la chaîne RAG, notamment récupération, ancrage, citations, réponses de repli, longueur et stabilité. Troisièmement, l'évaluation reste partiellement automatique et devrait être complétée par une analyse humaine plus systématique de la pertinence argumentative, de la complétude et de la justesse fine des réponses produites.

Les enjeux éthiques tiennent précisément à cette chaîne d'évaluation. Un système RAG peut produire une réponse plausible tout en s'appuyant sur des sources partielles, fragiles, mal sélectionnées ou insuffisamment attribuées. Ce risque de confiance déplacée est particulièrement important dans des contextes documentaires, scientifiques, juridiques, médicaux ou administratifs. La traçabilité constitue donc un enjeu éthique autant que méthodologique : rendre visibles les documents récupérés, les passages transmis au générateur, les contraintes de prompt, les citations produites et les réponses de repli permet de limiter l'opacité épistémique du système.

Cette traçabilité ne garantit toutefois pas la fiabilité du système. L'évaluation demeure dépendante du corpus, des requêtes, des métriques et de l'interprétation humaine des indicateurs. Les prolongements devront intégrer l'analyse des biais, la fiabilité des sources, la robustesse multilingue et les conditions sociales d'usage des systèmes RAG.

7 Conclusion

Cet article a présenté *Hermes-HSF* comme une plateforme expérimentale pour l'évaluation stratifiée et traçable des systèmes RAG. Il ne s'agissait pas de proposer une métrique unique, mais d'articuler des métriques classiques de recherche d'information avec des indicateurs opérationnels portant sur la récupération documentaire, le contexte, la génération, l'attribution aux sources, les réponses de repli et la stabilité.

Les résultats obtenus sur *SciFact* montrent que ces niveaux ne varient pas nécessairement dans le même sens. Les configurations fondées sur des unités textuelles avec reclassement améliorent les métriques de recherche documentaire, sans garantir une meilleure réponse générée. La configuration documentaire conserve le meilleur ancrage au contexte, tandis que la configuration contrainte accroît fortement la visibilité des citations sans progression proportionnelle de l'ancrage effectif.

Cette dissociation confirme l'intérêt d'une méthodologie stratifiée : une réponse RAG doit être rapportée à la chaîne de médiations qui la rend possible. Les prolongements porteront sur des corpus francophones et multilingues, des domaines plus hétérogènes, une évaluation humaine plus systématique et les enjeux éthiques liés à la traçabilité, aux sources et à la confiance.

Références

DENIS C. (2025). Vers une coconstruction harmonieuse du sens entre langage humain et machinique au sein de la fabrique sémiotique hybride. *Socio-anthropologie*, **52**, 37–50. DOI : 10.4000/15qda.

JÄRVELIN K. & KEKÄLÄINEN J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, **20**(4), 422–446. DOI : 10.1145/582415.582418.

KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

MANNING C. D., RAGHAVAN P. & SCHÜTZE H. (2008). *Introduction to Information Retrieval*. Cambridge : Cambridge University Press.

NOGUEIRA R. & CHO K. (2019). Passage re-ranking with BERT. Preprint arXiv.

THAKUR N., MÜLLER T., REIMERS N. & GUREVYCH I. (2025). MIRAGE-Bench : Automatic multilingual benchmark arena for language-agnostic RAG evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

THAKUR N., REIMERS N., DAXENBERGER J. & GUREVYCH I. (2021). BEIR : A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

WITTGENSTEIN L. (2014). *Recherches philosophiques*. Gallimard. Édition française.

YU H., GAN A., ZHANG K., TONG S., LIU Q. & LIU Z. (2024). Evaluation of retrieval-augmented generation : A survey. Preprint arXiv.