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

# Réévaluation de FACTUM : étude de réplication et analyse inter-modèles sur des modèles de langage open-weight

Hichem SEMMAR[1], Eric SANJUAN[2]

RÉSUMÉ

Cet article réévalue FACTUM, un cadre mécanistique pour la détection des hallucinations de citation dans les systèmes de génération augmentée par récupération à réponses longues. Nous répliquons et étendons son évaluation sur des modèles open-weight, en comparant LLaMA 3.2 3B, LLaMA 3.1 8B, Ministral 3B et Ministral 8B. Nous étudions à la fois des baselines avec juge externe et avec auto-évaluation du modèle, tout en clarifiant la convention FACTUM selon laquelle les citations incorrectes sont traitées comme classe positive par les métriques sklearn. Nos résultats montrent que les tendances FACTUM sont largement reproductibles sur LLaMA, tandis que Ministral reste exploitable mais obtient des scores légèrement inférieurs. Ces résultats suggèrent que les signaux de détection se transfèrent seulement partiellement entre familles de modèles et restent sensibles à la calibration, à la tokenisation, à l'architecture et aux choix d'implémentation. Nous discutons enfin la faisabilité d'utiliser CAS, BAS, PFS et PAS comme indicateurs de confiance pour des systèmes RAG industriels.

MOTS-CLÉS : modèles de langage, hallucination de citation, génération augmentée par récupération, interprétabilité, réplication.

ABSTRACT

**Revisiting FACTUM : A Replication and Cross-Model Evaluation on Open-Weight Language Models**

This paper revisits FACTUM, a mechanistic framework for detecting citation hallucinations in long-form retrieval-augmented generation. We replicate and extend its evaluation on open-weight models, comparing LLaMA 3.2 3B, LLaMA 3.1 8B, Ministral 3B, and Ministral 8B. We study both external-judge and self-judge confidence baselines, and clarify the FACTUM label convention, where incorrect citations are treated as the positive class for sklearn metrics. Our results show that FACTUM-style trends are largely reproducible on LLaMA models, while Ministral models remain usable but obtain slightly weaker scores. These findings suggest that citation-hallucination signals transfer only partially across model families and remain sensitive to calibration, tokenization, architecture, and implementation choices. We also discuss the feasibility of using CAS, BAS, PFS, and PAS as confidence indicators in industrial RAG systems.

KEYWORDS: language models, citation hallucination, retrieval-augmented generation, interpretability, replication.

## 1 Introduction

Les grands modèles de langage (LLM) sont de plus en plus utilisés pour la réponse à des questions, le résumé et la génération de textes longs. Dans la génération augmentée par récupération (RAG), la


Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

fiabilité factuelle est particulièrement importante : un modèle peut produire une réponse qui semble bien étayée tout en utilisant mal, en attribuant incorrectement ou en inventant des citations[2,3,10].

FACTUM[2] aborde ce problème en étudiant l'hallucination de citation à travers le comportement interne des modèles transformers. Il introduit des scores mécanistiques tels que le Contextual Alignment Score (CAS), le Beginning-of-Sentence Attention Score (BAS), le Parametric Force Score (PFS) et le Pathway Alignment Score (PAS), qui visent à capturer l'interaction entre les preuves contextuelles et la mémoire paramétrique lors de la génération des citations[2]. Plutôt que de traiter l'hallucination uniquement comme une erreur de récupération ou de sortie, FACTUM la formule comme un défaut de coordination entre ce que le modèle lit dans le contexte et ce qu'il rappelle depuis ses paramètres internes.

Dans ce travail, nous réexaminons FACTUM selon trois axes. Premièrement, nous répliquons l'évaluation sur des modèles LLaMA afin de vérifier si les tendances originales peuvent être retrouvées avec notre implémentation. Deuxièmement, nous étendons l'analyse aux modèles Ministral afin d'examiner si les signaux de type FACTUM se transfèrent entre familles de modèles. Troisièmement, nous clarifions le rôle des baselines de confiance et des conventions d'annotation, en particulier la distinction entre l'évaluation par un juge externe et l'utilisation de chaque modèle comme son propre juge.

Nos résultats suggèrent que les tendances de type FACTUM sont partiellement reproductibles sur les modèles LLaMA, mais moins stables lorsqu'elles sont transférées aux modèles Ministral. En particulier, nous observons une baisse de plusieurs indicateurs prédictifs, notamment l'AUC et le F1, ce qui peut refléter des différences propres aux modèles en matière de calibration, de tokenisation et de distribution de préentraînement. Ces résultats soulignent la nécessité d'évaluer les signaux mécanistiques de fiabilité sur plusieurs familles de modèles open-weight, plutôt que de supposer qu'ils se transfèrent sans changement.

Les contributions de cet article sont triples :

- nous proposons une réplication indépendante des tendances d'évaluation liées à FACTUM sur LLaMA 3.1 8B et LLaMA 3.2 3B ;
- nous étendons l'analyse aux modèles Ministral et rapportons des différences de performance entre modèles ;
- nous clarifions la manière dont le choix du modèle juge et les conventions d'annotation affectent l'interprétation des baselines de confiance et des métriques de classification.

# 2 FACTUM et hallucination de citation

FACTUM étudie l'hallucination de citation dans le RAG à réponses longues et propose quatre scores mécanistiques : CAS, BAS, PFS et PAS[2]. La méthode repose sur l'hypothèse selon laquelle un comportement de citation correct dépend de la coordination entre l'attention contextuelle et le rappel paramétrique, plutôt que d'un unique signal scalaire de confiance. L'article original rapporte que ces scores permettent de distinguer les comportements de citation corrects et incorrects et donnent un aperçu de la circulation de l'information à travers les couches du transformer.

**Parametric Force Score (PFS)** Mesure l'intensité avec laquelle le réseau feed-forward (FFN), interprétable comme la mémoire paramétrique interne du modèle, modifie la représentation d'un token pendant le passage avant à une couche donnée. Un PFS élevé indique que le FFN modifie fortement l'embedding du token, ce qui suggère que le modèle s'appuie davantage sur des connaissances

stockées en interne que sur l'information contextuelle. Cette métrique aide donc à identifier les couches du transformer responsables des plus grands changements de représentation pendant le traitement.

$$\mathrm{PFS}^{(l)} = \left\|\mathbf{v}_{\text{ffn}}^{(l)}\right\|_2 = \sqrt{\sum_{k=1}^{d}\left(v_k^{(l)}\right)^2} \tag{1}$$

$$\mathbf{v}_{\text{ffn}}^{(l)} = \mathbf{x}_{\text{post-ffn}}^{(l)} - \mathbf{x}_{\text{pre-ffn}}^{(l)} \tag{2}$$

**Pathway Alignment Score (PAS)** Évalue si les deux principaux mécanismes de mise à jour d'une couche transformer — la mise à jour par attention, qui incorpore l'information contextuelle issue des autres tokens, et la mise à jour par FFN, qui injecte de la connaissance paramétrique — déplacent la représentation du token dans une direction similaire de l'espace de représentation. Lorsque ces deux voies sont alignées, l'information contextuelle et la connaissance interne se renforcent mutuellement. Le PAS renseigne donc sur la quantité d'information provenant d'autres tokens, y compris de documents externes dans les systèmes fondés sur la récupération, intégrée dans la représentation du token à une couche donnée.

$$\mathrm{PAS}^{(l)} = \frac{\mathbf{v}_{\text{attn}}^{(l)} \cdot \mathbf{v}_{\text{ffn}}^{(l)}}{\left\|\mathbf{v}_{\text{attn}}^{(l)}\right\|_2 \cdot \left\|\mathbf{v}_{\text{ffn}}^{(l)}\right\|_2} \tag{3}$$

$$\mathbf{v}_{\text{attn}}^{(l)} = \mathbf{x}_{\text{pre-ffn}}^{(l)} - \mathbf{x}_{\text{input}}^{(l)} \tag{4}$$

**Contextual Alignment Score (CAS)** Mesure dans quelle mesure la représentation finale du token s'aligne avec les représentations des documents sources auxquels le modèle a prêté attention. Autrement dit, il évalue à quel point le token généré reste ancré dans les preuves récupérées. Un CAS élevé suggère que la représentation du token ressemble sémantiquement au contenu des documents de support, ce qui indique que la sortie du modèle est cohérente avec l'information externe consultée pendant la génération.

$$\mathrm{CAS}^{(h,l)}(t_i) = \frac{\left(\sum_{t_j \in \mathcal{T}_C} A_{i,j}^{(h,l)}\, \mathbf{h}_{t_j}^{(L)}\right) \cdot \mathbf{h}_{t_i}^{(L)}}{\left\|\sum_{t_j \in \mathcal{T}_C} A_{i,j}^{(h,l)}\, \mathbf{h}_{t_j}^{(L)}\right\|_2 \cdot \left\|\mathbf{h}_{t_i}^{(L)}\right\|_2} \tag{5}$$

**Beginning-of-Sentence Attention Score (BAS)** Quantifie l'attention qu'un modèle alloue au token spécial de début de séquence ($\langle \mathrm{s} \rangle$) à une couche donnée. Une attention élevée portée à ce token peut indiquer que le modèle s'appuie sur une forme de synthèse interne ou sur un mécanisme de résumé global. Le BAS renseigne donc sur la mesure dans laquelle le modèle agrège l'information via le token de début de séquence lors de la formation de ses représentations internes.

$$\mathrm{BAS}^{(l,h)}(t_i) = A_{i,1}^{(l,h)} \tag{6}$$

Le présent travail ne cherche pas à modifier le cadre FACTUM lui-même. Il évalue plutôt la reproductibilité des tendances rapportées et examine si les signaux prédictifs associés restent stables entre familles de modèles.

# 3 Méthodologie

Ce travail étend FACTUM selon trois axes principaux. Premièrement, nous évaluons un ensemble plus large de modèles open-weight afin de tester si les signaux d'hallucination de citation de type FACTUM se transfèrent au-delà du cadre original. Deuxièmement, nous comparons des baselines avec juge externe à un cadre d'auto-évaluation, où chaque modèle est évalué à partir de ses propres logits. Troisièmement, nous clarifions la convention d'annotation utilisée pour définir les citations correctes et incorrectes, car elle affecte l'interprétation de la précision, du rappel, du F1 et des résultats de matrice de confusion.

## 3.1 Modèles

Nous évaluons les LLM open-weight suivants :

— LLaMA 3.2 3B et LLaMA 3.1 8B, utilisés comme principales cibles de réplication ;
— Ministral 3B et Ministral 8B, utilisés pour la comparaison inter-modèles.

Le choix de ces modèles reflète deux objectifs complémentaires. Les modèles LLaMA fournissent un test direct de reproductibilité par rapport au cadre FACTUM original. Les modèles Ministral permettent au contraire d'évaluer la transférabilité vers une autre famille de modèles open-weight, avec des caractéristiques distinctes de préentraînement et de déploiement.

## 3.2 Cadre du modèle juge

Pour les baselines fondées sur la confiance, nous comparons deux cadres. Dans le cadre avec juge externe, un modèle séparé calcule les scores de confiance. Dans le cadre d'auto-évaluation, chaque modèle est évalué à partir de ses propres logits, de sorte que le score reflète la confiance du modèle ayant généré la réponse plutôt que l'incertitude d'un autre modèle.

## 3.3 Convention d'annotation

FACTUM formule la détection d'hallucination de citation comme une classification binaire sur les tokens de citation :

$$0 = \text{good/correct}, \qquad 1 = \text{bad/incorrect}.$$

Comme $\mathrm{sklearn}$ traite par défaut le label $1$ comme classe positive, la précision, le rappel et le F1 de FACTUM mesurent la détection des citations incorrectes. Nous rendons cette convention explicite avant de rapporter les résultats.

# 4 Métriques d'évaluation

## 4.1 Jeu de données

À la suite de FACTUM, nous évaluons la détection d'hallucination de citation dans un cadre RAG multi-document à réponses longues, en utilisant le jeu de données TREC NeuCLIR 2024 de génération de rapports[2,7,17]. Dans cette tâche, un modèle reçoit une requête et un ensemble de documents sources récupérés, puis génère un rapport en anglais où chaque phrase factuelle doit se terminer par une citation en ligne de la forme $[\mathrm{Source}: \ \#]$. FACTUM formule la détection d'hallucination de citation

comme un problème de classification binaire au niveau du token, sur le chiffre de citation lui-même, puisque ce token établit le lien entre l'assertion générée et la source référencée[2]. Une citation est considérée comme correcte lorsque la source citée soutient l'assertion associée, et comme hallucinée lorsque la source référencée n'atteste pas cette assertion[2].

Les expériences FACTUM originales utilisent 15 documents récupérés en entrée du générateur. Bien que NeuCLIR soit un benchmark cross-lingual, FACTUM mène l'analyse dans un cadre monolingue anglais en utilisant des versions anglaises traduites automatiquement des documents sources et en générant des rapports en anglais. Comme la trace de génération propre au modèle est nécessaire à l'analyse mécanistique, les labels sont attribués directement aux sorties générées. L'article FACTUM utilise une procédure automatisée de type LLM-as-a-judge fondée sur Llama-3.1-70B-Instruct, limitée à des jugements d'attestation étroits inspirés du cadre d'évaluation de rapports ARGUE[12]. Les auteurs valident ces labels au moyen de contrôles de stabilité et d'annotations humaines, et rapportent un accord substantiel avec les jugements humains.

Dans notre implémentation actuelle, nous utilisons les fichiers NeuCLIR24 pour $\mathrm{ministral\text{-}3\text{-}3b\text{-}base\text{-}2512}$. Le fichier de réponses contient les rapports générés et les labels au niveau des citations, tandis que le fichier d'information sur les sources stocke le prompt correspondant, les passages récupérés et le contexte source. Le code associe $\mathrm{label\_type} = \mathrm{good}$ à la classe $0$, correspondant à une citation correcte, et tous les labels non $\mathrm{good}$ à la classe $1$, correspondant à une citation incorrecte ou hallucinée. Dans les données Ministral locales actuelles, on compte 38 réponses générées, 36 réponses avec labels et 949 segments de citation annotés au total : 648 annotés $\mathrm{good}$ et 301 annotés $\mathrm{bad}$.

## 4.2 Métriques de baseline

En plus de la motivation mécanistique de FACTUM, la comparaison empirique repose sur quatre signaux probabilistes extraits du comportement des modèles afin d'analyser leur confiance lors de la génération des citations :

### Perplexity

La perplexité[4] mesure la vraisemblance d'une séquence selon le modèle :

$$\mathrm{PPL} = \exp\left(-\frac{1}{V}\sum_{i=1}^{V}\log p\left(\mathrm{prompt}, y_i \mid y_{<i}\right)\right)$$

### Entropy

L'entropie[16] mesure l'incertitude de la distribution du prochain token du modèle à l'étape $t$ :

$$H_t = -\frac{1}{V}\sum_{i=1}^{V} p_i \log p_i$$

### Energy Score

Le score d'énergie[9] est calculé directement à partir des logits du modèle :

$$E_t = -\log \sum_{i=1}^{V} \exp\left(z_i^{(v)}\right)$$

## P(True) Score

$P(\text{True})$[5] estime la probabilité auto-évaluée par le modèle qu'une réponse générée soit correcte :

$$P(\text{True}) = p(\text{“true”} \mid \text{q}, \text{a}, \text{p})$$

## 4.3 Métriques et annotations de classes

Dans FACTUM, l'évaluation est binaire, mais la convention de labels diffère de celle utilisée dans la formulation de référence. Les annotations du jeu de données sont associées comme suit :

$$0 = \text{good/correct}, \qquad 1 = \text{bad/incorrect}.$$

Ainsi, lorsque FACTUM calcule la précision, le rappel et le F1 avec **sklearn**, la classe positive n'est pas la classe des réponses correctes. Par défaut, **sklearn** traite le label $1$ comme la classe positive. Par conséquent, la précision, le rappel et le F1 rapportés par FACTUM évaluent principalement la détection des réponses incorrectes, plutôt que l'acceptation des réponses correctes.

Par conséquent, le rappel de FACTUM est le rappel des réponses incorrectes :

$$\text{Recall} = \frac{\text{réponses incorrectes correctement détectées comme incorrectes}}{\text{toutes les réponses réellement incorrectes}}.$$

Il ne doit pas être décrit comme un rappel des réponses correctes, sauf si les labels sont inversés ou si $\text{pos\_label} = 0$ est explicitement utilisé.

Pour les baselines à score faible, à savoir Perplexity, LN-Entropy et Energy Score, FACTUM applique la règle de seuillage suivante :

$$\hat{y}_i = \begin{cases} 1, & \text{si } s_i > \tau, \\ 0, & \text{si } s_i \leq \tau. \end{cases}$$

Ainsi, les scores élevés prédisent la classe bad/incorrect, tandis que les scores faibles prédisent la classe good/correct.

Pour la baseline à score élevé, $P(\text{True})$ Query, la direction du score est inversée :

$$\hat{y}_i = \begin{cases} 1, & \text{si } s_i < 0.5, \\ 0, & \text{si } s_i \geq 0.5. \end{cases}$$

Ainsi, les valeurs élevées de $P(\text{True})$ prédisent la classe good/correct, tandis que les valeurs faibles prédisent la classe bad/incorrect. Cependant, la dénomination de la matrice de confusion reste centrée sur le label $1$, car $\text{sklearn}$ traite $1 = \text{bad/incorrect}$ comme la classe positive.

C'est la différence principale avec les formulations où la classe positive correspond aux réponses correctes. Dans FACTUM, puisque bad/incorrect est encodé par $1$, $\text{sklearn}$ traite les réponses incorrectes comme la classe positive. Par conséquent, un cas tel que

$$\text{réponse incorrecte} + \text{perplexité élevée} \rightarrow \text{prédiction incorrecte}$$

| Vérité terrain | Score baseline | Classe prédite | Résultat FACTUM | Notre résultat |
|---|---|---|---|---|
| Correct/good 0 | Faible | Correct/good 0 | TN | TP |
| Correct/good 0 | Élevé | Incorrect/bad 1 | FP | FN |
| Incorrect/bad 1 | Faible | Correct/good 0 | FN | FP |
| Incorrect/bad 1 | Élevé | Incorrect/bad 1 | TP | TN |

TABLE 1 – Résultats de matrice de confusion pour les baselines à score faible. FACTUM suit la convention $\mathrm{sklearn}$, où $1 =$ bad/incorrect est la classe positive, tandis que notre interprétation traite les citations correctes comme classe positive.

| Vérité terrain | Score $P(\text{True})$ | Classe prédite | Résultat FACTUM | Notre résultat |
|---|---|---|---|---|
| Correct/good 0 | Élevé | Correct/good 0 | TN | TP |
| Correct/good 0 | Faible | Incorrect/bad 1 | FP | FN |
| Incorrect/bad 1 | Élevé | Correct/good 0 | FN | FP |
| Incorrect/bad 1 | Faible | Incorrect/bad 1 | TP | TN |

TABLE 2 – Résultats de matrice de confusion pour $P(\text{True})$ Query. FACTUM suit la convention $\mathrm{sklearn}$, où $1 =$ bad/incorrect est la classe positive, tandis que notre interprétation traite les citations correctes comme classe positive.

est compté comme un vrai positif dans FACTUM, alors qu'il serait appelé vrai négatif dans une convention où les réponses incorrectes constituent la classe négative.

# 5 Résultats

Nous évaluons plusieurs baselines fondées sur la confiance, notamment Perplexity, LN-Entropy, Energy Score et $P(\text{True})$. Ces scores sont dérivés des logits du modèle et mesurent donc à quel point le token de citation généré, ou l'énoncé jugé, est probable, incertain ou énergétiquement favorisé pour un modèle de langage donné[5,11,9]. Un choix méthodologique central concerne l'identité du modèle utilisé pour calculer ces scores. Dans un cadre de type FACTUM avec juge externe, les scores de confiance sont calculés avec un modèle juge séparé, ici Llama-2-7B. Toutefois, ce cadre ne mesure pas nécessairement la confiance du modèle qui a produit la réponse ; il mesure plutôt l'incertitude du juge externe vis-à-vis de la sortie d'un autre modèle.

Pour cette raison, nous rapportons également un cadre d'auto-évaluation, dans lequel chaque modèle générateur est évalué à partir de ses propres logits. Ce cadre est mieux aligné avec l'interprétation de ces baselines comme mesures de confiance interne : la perplexité reflète à quel point le token de citation généré est surprenant pour le modèle générateur lui-même, l'entropie reflète la netteté de la distribution du prochain token de ce modèle, et l'énergie reflète la force de ses propres activations de logits. Utiliser le générateur comme son propre juge évite donc de confondre la correction des citations avec un décalage entre modèles, des différences de calibration ou des différences de tokenisation et de distributions de probabilité apprises. Nous rapportons deux ensembles de résultats : l'un utilisant le juge externe Llama-2-7B, conformément à la comparaison avec juge externe, et l'autre utilisant

chaque modèle comme son propre juge.

De plus, bien que les mêmes métriques FACTUM soient évaluées pour LLaMA et Ministral, leur calcul diffère au niveau de l'implémentation. Dans le cadre LLaMA, le code transformer FACTUM modifié expose directement les états internes des voies nécessaires à POS/PAS et PFS, comme le flux résiduel avant l'attention, après l'attention et après le bloc FFN. Dans Ministral, ces états spécifiques à FACTUM ne sont pas fournis par défaut : le modèle expose des sorties standard telles que les logits, les états cachés, les attentions et les valeurs de cache, mais pas la décomposition complète des voies. Nous reconstruisons donc manuellement les quantités POS/PAS manquantes en capturant les activations intermédiaires pertinentes pendant le passage avant. En outre, le checkpoint Ministral est chargé via une interface de modèle multimodal, alors que notre tâche est purement textuelle ; par conséquent, les expériences sont restreintes au composant de modèle de langage sous-jacent afin de garder la comparaison avec LLaMA centrée sur la génération des tokens de citation.

**Évalué par LLaMA 2 7B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.4785 | 0.0050 | 0.3627 | 0.1260 | 0.6563 | 0.1901 |
| LN-Entropy | 0.5378 | 0.0636 | 0.7018 | 0.2278 | 0.2340 | 0.1489 |
| Energy Score | 0.5338 | 0.0624 | 0.5237 | 0.1056 | 0.3952 | 0.1474 |
| P(True) Query | 0.4714 | -0.0307 | 0.1603 | 0.1603 | 0.9240 | 0.2711 |
| FACTUM | 0.5287 | 0.0417 | 0.5553 | 0.2014 | 0.4465 | 0.2566 |

**Évalué par LLaMA 3.1 8B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.6008 | 0.0418 | 0.6403 | 0.2118 | 0.4784 | 0.2823 |
| LN-Entropy | 0.6082 | 0.1519 | 0.6262 | 0.2071 | 0.4862 | 0.2750 |
| Energy Score | 0.5542 | 0.0625 | 0.4666 | 0.1721 | 0.6328 | 0.2529 |
| P(True) Query | 0.4894 | -0.0039 | 0.2245 | 0.1589 | 0.9241 | 0.2669 |
| FACTUM | 0.5287 | 0.0417 | 0.5553 | 0.2014 | 0.4465 | 0.2566 |

TABLE 3 – Résultats finaux en validation croisée sur LLaMA 3.1 selon deux cadres d'évaluation.

**Évalué par LLaMA 2 7B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.6366 | 0.0850 | 0.5634 | 0.3749 | 0.6695 | 0.4155 |
| LN-Entropy | 0.6429 | 0.1943 | 0.4932 | 0.3295 | 0.6888 | 0.3728 |
| Energy Score | 0.5517 | 0.0634 | 0.4884 | 0.2936 | 0.7031 | 0.3895 |
| P(True) Query | 0.5455 | 0.0701 | 0.2986 | 0.2986 | 0.9560 | 0.4524 |
| FACTUM | 0.5709 | 0.1076 | 0.5209 | 0.3459 | 0.5747 | 0.4017 |

**Évalué par LLaMA 3.2 3B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.5163 | -0.0130 | 0.4460 | 0.3148 | 0.7399 | 0.4290 |
| LN-Entropy | 0.5138 | 0.0168 | 0.4058 | 0.3079 | 0.8114 | 0.4367 |
| Energy Score | 0.4468 | -0.0840 | 0.5808 | 0.3152 | 0.2829 | 0.2014 |
| P(True) Query | 0.4721 | -0.0453 | 0.3169 | 0.2972 | 0.9557 | 0.4480 |
| FACTUM | 0.5709 | 0.1076 | 0.5209 | 0.3459 | 0.5747 | 0.4017 |

TABLE 4 – Résultats finaux en validation croisée sur LLaMA 3.2 selon deux cadres d'évaluation.

**Évalué par LLaMA 2 7B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.6366 | 0.0850 | 0.5634 | 0.3749 | 0.6695 | 0.4155 |
| LN-Entropy | 0.6429 | 0.1943 | 0.4932 | 0.3295 | 0.6888 | 0.3728 |
| Energy Score | 0.5517 | 0.0634 | 0.4884 | 0.2936 | 0.7031 | 0.3895 |
| P(True) Query | 0.5455 | 0.0701 | 0.2986 | 0.2986 | 0.4440 | 0.4524 |
| FACTUM | 0.4982 | -0.0156 | 0.4991 | 0.2703 | 0.4634 | 0.3131 |

**Évalué par Ministral 3B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.5752 | 0.0291 | 0.5247 | 0.3248 | 0.5497 | 0.3837 |
| LN-Entropy | 0.5108 | 0.0367 | 0.5767 | 0.2985 | 0.3551 | 0.2941 |
| Energy Score | 0.5294 | 0.0506 | 0.4826 | 0.2896 | 0.5526 | 0.3356 |
| P(True) Query | 0.4736 | -0.0287 | 0.4748 | 0.2737 | 0.4438 | 0.3242 |
| FACTUM | 0.4982 | -0.0156 | 0.4991 | 0.2703 | 0.4634 | 0.3131 |

TABLE 5 – Résultats finaux en validation croisée sur Ministral 3B selon deux cadres d'évaluation.

**Évalué par LLaMA 2 7B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.6366 | 0.0850 | 0.5634 | 0.3749 | 0.6695 | 0.4155 |
| LN-Entropy | 0.6429 | 0.1943 | 0.4932 | 0.3295 | 0.6888 | 0.3728 |
| Energy Score | 0.5517 | 0.0634 | 0.4884 | 0.2936 | 0.7031 | 0.3895 |
| P(True) Query | 0.5455 | 0.0701 | 0.2986 | 0.2986 | 0.7000 | 0.4524 |
| FACTUM | 0.4774 | -0.0552 | 0.4302 | 0.2851 | 0.6145 | 0.3655 |

**Évalué par Ministral 8B**

| Baseline | AUC | PCC | Acc. | Préc. | Rapp. | F1 |
|---|---|---|---|---|---|---|
| Perplexity | 0.5539 | 0.0055 | 0.4951 | 0.3269 | 0.7035 | 0.4381 |
| LN-Entropy | 0.5311 | 0.0549 | 0.5531 | 0.3308 | 0.4633 | 0.3614 |
| Energy Score | 0.5375 | 0.0580 | 0.5113 | 0.3107 | 0.5036 | 0.3548 |
| P(True) Query | 0.5002 | 0.0102 | 0.4423 | 0.3091 | 0.7001 | 0.4181 |
| FACTUM | 0.4774 | -0.0552 | 0.4302 | 0.2851 | 0.6145 | 0.3655 |

TABLE 6 – Résultats finaux en validation croisée sur Ministral 8B selon deux cadres d'évaluation.

# 6 Contribution industrielle

D'un point de vue industriel, cette expérimentation étudie si les signaux de confiance internes peuvent être utilisés comme indicateurs opérationnels de fiabilité dans les systèmes RAG à réponses longues. Les baselines classiques comme la perplexité, l'entropie, le score d'énergie et $P(\mathrm{True})$ estiment le degré de confiance d'un modèle dans sa propre réponse générée. Ce point est important en production, car les systèmes RAG sont souvent déployés pour réduire l'hallucination, mais la récupération seule ne garantit pas que la réponse générée soit correctement ancrée dans les preuves récupérées[8,15]. Dans ce contexte, les métriques de confiance peuvent soutenir le monitoring, le triage et le contrôle des risques en aval, en identifiant les réponses pour lesquelles le modèle semble incertain, ou les cas où le modèle est très confiant malgré une citation incorrecte.

FACTUM prolonge cette idée en proposant des métriques mécanistiques qui ne se limitent pas à scorer la probabilité de surface de la réponse, mais examinent aussi la manière dont le modèle coordonne en interne les preuves contextuelles et la connaissance paramétrique pendant la génération des citations[2]. Dans notre expérimentation, nous évaluons donc la faisabilité technique d'utiliser les métriques FACTUM, à savoir CAS, BAS, PFS et PAS, comme indicateurs de la confiance du modèle dans sa décision de citation. Une prédiction peut être considérée comme fiable lorsque la métrique attribue une confiance élevée à une citation effectivement correcte, ou une confiance faible à une citation qui s'avère ensuite fausse. À l'inverse, une confiance élevée dans une réponse incorrecte est particulièrement pertinente pour le déploiement industriel, car elle correspond à un mode de défaillance dans lequel le système peut présenter une information non étayée comme fiable.

Par rapport à l'étude FACTUM originale, centrée sur les modèles de la famille LLaMA, nos expériences évaluent également les modèles Ministral. Cet aspect est industriellement pertinent dans le contexte de déploiement français et européen, où les modèles Mistral sont attractifs en raison de leur disponibilité en open-weight et de leur compatibilité avec des déploiements locaux ou cloud[13,14]. Nos résultats suggèrent que les modèles Ministral restent utilisables avec des métriques de type FACTUM, même si leurs scores sont légèrement inférieurs à ceux obtenus avec les modèles LLaMA. Cela indique que l'approche est techniquement transférable, mais que les effets de calibration et d'architecture propres aux modèles doivent être pris en compte avant de déployer ces métriques comme signaux de confiance en production.

Enfin, ces mesures de confiance sont également utiles au-delà de la détection d'hallucinations. Dans des applications réglementées ou juridiquement sensibles, les réponses à la fois correctes et associées à une confiance élevée peuvent nécessiter une surveillance, en particulier lorsque le modèle reproduit un texte proche d'une source protégée. Des travaux antérieurs ont montré que les modèles de langage peuvent mémoriser et parfois reproduire mot à mot des données d'entraînement[1], et des études centrées sur le droit d'auteur ont examiné le risque de redistribution de textes protégés via les sorties des modèles[6]. Ainsi, les réponses correctes à forte confiance ne sont pas seulement utiles pour l'évaluation de la fiabilité ; elles peuvent aussi aider à identifier les cas où la sortie du système devrait être auditée en matière de provenance, d'attribution et de conformité avec la propriété intellectuelle.

# Références

[1] CARLINI N., TRAMÈR F., WALLACE E., JAGIELSKI M., HERBERT-VOSS A., LEE K., ROBERTS A., BROWN T., SONG D., ERLINGSSON U., OPREA A. & RAFFEL C. (2021). Extracting training data from large language models. In *USENIX Security Symposium*.

[2] DASSEN M., KOTULA R., MURRAY K., YATES A., LAWRIE D., KAYI E., MAYFIELD J. & DUH K. (2026). Factum : Mechanistic detection of citation hallucination in long-form rag. *arXiv preprint arXiv :2601.05866*.

[3] DING Y., FACCIANI M., JOYCE E., POUDEL A., BHATTACHARYA S., VEERAMANI B., AGUINAGA S. & WENINGER T. (2025). Citations and trust in llm generated responses. *arXiv preprint arXiv*.

[4] HUANG X. *et al.* (2025). Reppl : Recalibrating perplexity for large language models. *arXiv preprint arXiv :2505.15386*.

[5] KADAVATH S., CONERLY T., ASKELL A., HENIGHAN T., DRAIN D., PEREZ E. *et al.* (2022). Language models (mostly) know what they know. *arXiv preprint arXiv :2207.05221*.

[6] KARAMOLEGKOU A., LI J., ZHOU L. & SØGAARD A. (2023). Copyright violations and large language models. In *Proceedings of EMNLP*.

[7] LAWRIE D., MACAVANEY S., MAYFIELD J., MCNAMEE P., OARD D. W., SOLDAINI L. & YANG E. (2025). Overview of the TREC 2024 NeuCLIR track.

[8] LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, **33**, 9459–9474.

[9] LIU W., WANG X., OWENS J. D. & LI Y. (2020). Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[10] MAGESH V., SURANI F., DAHL M., SUZGUN M., MANNING C. D. & HO D. E. (2025). Hallucination-free ? assessing the reliability of leading ai legal research tools.

[11] MALININ A. & GALES M. (2021). Uncertainty estimation in autoregressive structured prediction.

[12] MAYFIELD J., YANG E., LAWRIE D., MACAVANEY S., MCNAMEE P., OARD D. W., SOLDAINI L., SOBOROFF I., WELLER O., KAYI E., SANDERS K., MASON M. & HIBBLER N. (2024). On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1904–1915 : Association for Computing Machinery. DOI : 10.1145/3626772.3657846.

[13] MISTRAL AI (2026a). Mistral ai documentation : Deployment. https://docs.mistral.ai/models/deployment.

[14] MISTRAL AI (2026b). Mistral ai models. https://mistral.ai/models.

[15] NIU C., WU J., ZHANG Y. *et al.* (2024). Ragtruth : A hallucination corpus for developing trustworthy retrieval-augmented language models.

[16] SCHARRINGHAUSEN M. *et al.* (2026). Entropy in large language models. *arXiv preprint arXiv :2602.20052*.

[17] TREC NEUCLIR ORGANIZERS (2024). TREC NeuCLIR track. https://neuclir.github.io/. Accessed 2026-05-15.