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

Analyse des signaux de mémorisation dans les modèles de langage à poids ouverts

Hichem Semmar Eric SanJuan
Laboratoire d'Informatique d'Avignon (LIA) – SFR Agorantic
Avignon Université, France
{hichem.semmar, eric.sanjuan}@univ-avignon.fr

RÉSUMÉ

La mémorisation dans les grands modèles de langage (LLMs) soulève des enjeux importants dans les domaines sensibles impliquant la propriété intellectuelle (IP) ou des données personnelles, où la fiabilité et la traçabilité des contenus générés sont essentielles. Il demeure toutefois difficile de déterminer si une réponse résulte d'une généralisation ou d'une exposition préalable aux données d'entraînement sans accès direct à celles-ci. Dans ce travail, on étudie si des signaux internes permettent de caractériser un comportement de type mémorisation dans des modèles à poids ouverts. En nous appuyant sur FACTUM, on analyse des métriques internes telles que PFS et PAS afin d'étudier comment les modèles équilibrent mémoire paramétrique et information contextuelle. À l'aide d'un benchmark dérivé de la littérature COVID-19 soumise à des restrictions IP, on évalue des modèles Llama 3.1 8B, Llama 3.2 3B et Ministral 3-3B à travers des métriques de confiance classiques, notamment la perplexité, l'entropie, l'énergie et P(True). Les résultats montrent que les signaux internes et probabilistes corrèlent avec la justesse des réponses avec des implications pour l'audit des modèles et l'évaluation juridique.

MOTS-CLÉS : Grands modèles linguistiques, détection des hallucinations, mémorisation, exposition des données d'entraînement, interprétabilité des modèles, estimation de l'incertitude, perplexité, entropie, FACTUM, mémoire paramétrique, réglage fin, propriété intellectuelle.

ABSTRACT

Analysis of Memorization Signals in Open-Weight Language Models

Memorization in large language models (LLMs) raises important concerns in sensitive domains involving intellectual property (IP) and personal data, where reliability and traceability are essential. However, it remains unclear whether a model's behavior can reveal evidence of exposure to training data without direct access to the data itself. In this work, we investigate whether internal signals can characterize memorization-like behavior in open-weight models. Building on FACTUM, we analyze internal metrics such as PFS and PAS to study how models balance parametric memory and contextual information. Using a benchmark derived from COVID-19 scientific literature subject to IP restrictions, we evaluate Llama 3.1 8B, Llama 3.2 3B, and Ministral 3-3B through classical confidence metrics, including perplexity, entropy, energy, and P(True). Results show that internal and probabilistic signals correlate with answer correctness, with implications for model auditing and legal assessment.

KEYWORDS: Large Language Models, Hallucination Detection, Memorization, Training Data Exposure, Model Interpretability, Uncertainty Estimation, Perplexity, Entropy, FACTUM, Parametric Memory, Fine-Tuning, Intellectual Property.

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 1–12.

1 Introduction

Les grands modèles de langage (LLMs) sont de plus en plus utilisés dans des tâches à forte intensité de connaissances, telles que la réponse à des questions, le résumé et la génération de textes longs. Malgré leurs fortes capacités génératives, ces modèles restent vulnérables à l'incertitude factuelle, à l'hallucination et à la génération de contenus non traçables. Des travaux antérieurs ont étudié si les modèles peuvent reconnaître des références hallucinées [1], si les états internes peuvent révéler un risque d'hallucination [8], et comment la qualité des citations affecte la confiance dans les réponses générées [5]. D'autres approches ont proposé des pipelines dédiés à la détection des hallucinations pour les systèmes de génération augmentée par récupération [10].

Alors que la détection des hallucinations vise souvent à identifier des sorties non étayées ou fausses, le présent travail se concentre sur un problème lié mais distinct : la **mémorisation**. Un modèle peut produire une réponse parce qu'il l'a inférée à partir du contexte, généralisée à partir de connaissances connexes, ou parce qu'il a rencontré des informations similaires pendant l'entraînement ou le fine-tuning. Distinguer ces cas est difficile, en particulier lorsque les données d'entraînement originales ne sont pas disponibles. Cette difficulté devient particulièrement importante lorsque les sorties générées ressemblent à des contenus protégés ou soumis au droit d'auteur, car la sortie seule ne révèle pas si le modèle a mémorisé le contenu ou s'il a simplement généré une continuation plausible.

Cette question est particulièrement pertinente pour les **modèles à poids ouverts**. Contrairement aux modèles fermés, les modèles à poids ouverts donnent accès à l'architecture et aux paramètres appris, ce qui permet d'inspecter non seulement le texte final, mais aussi les probabilités des tokens, les logits, les représentations cachées et les activations couche par couche. Cette transparence ne révèle pas le jeu de données d'entraînement lui-même, mais elle permet d'étudier des régularités comportementales mesurables pouvant indiquer une familiarité de type mémorisation.

Dans ce travail, nous étudions si un comportement de type mémorisation peut être approché au moyen de signaux de confiance et de représentations internes. Des baselines classiques telles que la perplexité, l'entropie, l'energy score et $P(\mathrm{True})$ estiment dans quelle mesure un modèle soutient sa propre réponse. Nous étudions également des métriques internes de type FACTUM [4], en particulier PFS et PAS, qui fournissent des informations par token et par couche sur le calcul interne du modèle pendant la génération. Ces métriques diffèrent des approches générales de détection des hallucinations, telles qu'EigenScore [3] ou les scores fondés sur la véracité [12], qui évaluent principalement la fiabilité factuelle ou la cohérence sémantique plutôt que les schémas de mémorisation.

Notre objectif n'est donc pas seulement de détecter si une réponse est correcte ou incorrecte, mais d'examiner si des réponses correctes produites avec une forte confiance révèlent des traces internes plus marquées de mémorisation. Une réponse correcte générée avec une faible incertitude et un fort alignement interne peut indiquer que des informations similaires étaient fortement représentées pendant l'entraînement ou le fine-tuning. À l'inverse, une réponse incertaine ou instable peut suggérer une mémorisation plus faible ou une dépendance plus importante à la génération sous incertitude.

Ce problème possède également une dimension juridique pratique. Les tribunaux et les régulateurs peuvent devoir déterminer si les développeurs de LLMs ont utilisé des contenus protégés ou soumis au droit d'auteur pendant l'entraînement, alors que l'accès direct aux jeux de données d'entraînement est souvent indisponible. Des litiges récents, tels que *GEMA v. OpenAI*, illustrent la difficulté d'interpréter des sorties générées qui ressemblent fortement à des œuvres protégées [14]. Même lorsqu'une similarité textuelle est observée, la distinction entre mémorisation, généralisation et coïncidence reste techniquement difficile à établir. Dans cette affaire, OpenAI a soutenu que le contenu reproduit pouvait s'expliquer par une coïncidence hallucinée ou par une conséquence de la généralisation du modèle, tandis que la procédure juridique a traité les sorties comme une preuve de mémorisation.

Cela illustre le défi plus large consistant à utiliser le seul texte généré pour inférer si un contenu protégé était présent dans les données d'entraînement.

La contribution de ce travail consiste à explorer si des métriques fondées sur la confiance et des métriques internes de type FACTUM peuvent soutenir l'analyse de la mémorisation dans les modèles à poids ouverts. Plutôt que de traiter les sorties du modèle comme des preuves isolées, nous utilisons les probabilités et les activations internes propres au modèle afin d'étudier si certaines réponses présentent des signes plus forts de familiarité apprise. Cela constitue une première étape vers un cadre technique permettant de distinguer hallucination, génération incertaine et mémorisation possible.

2 Travaux connexes et métriques

Les travaux antérieurs sur la détection des hallucinations ont examiné à la fois le comportement des sorties et les états internes des modèles. Les approches fondées sur les sorties évaluent si le texte généré est factuellement fiable, sémantiquement cohérent ou véridiquement étayé [1,2,3]. D'autres travaux étudient si la confiance du modèle peut être inférée à partir des probabilités des tokens, de l'entropie ou de signaux probabilistes connexes [20]. Ces approches sont utiles pour détecter une génération peu fiable, mais elles n'abordent pas directement la question de savoir si la réponse d'un modèle reflète une familiarité de type mémorisation avec des informations vues pendant l'entraînement.

FACTUM a été initialement proposé pour détecter les hallucinations de citations dans des contextes RAG en génération longue [4]. Il analyse l'interaction entre le contexte externe et la mémoire paramétrique interne pendant la génération de citations, au moyen de plusieurs scores mécanistiques, notamment CAS, BAS, PFS et PAS. Dans ce travail, nous réutilisons ce cadre sous un autre angle. Plutôt que de traiter FACTUM principalement comme un détecteur d'hallucinations de citations, nous nous concentrons sur PFS et PAS comme signaux internes susceptibles de refléter la confiance du modèle et un comportement de type mémorisation. Ce choix est motivé par notre intérêt pour les modèles à poids ouverts, où les activations internes et les calculs couche par couche peuvent être inspectés directement.

Parallèlement à ces scores mécanistiques, nous évaluons des baselines classiques fondées sur la confiance : perplexité, entropie, energy score et $P(\mathrm{True})$. Ces métriques estiment, selon différentes perspectives, dans quelle mesure le modèle soutient sa propre réponse : vraisemblance de la séquence, incertitude de la distribution du prochain token, énergie des logits et auto-évaluation explicite. Ensemble, ces baselines et les métriques de type FACTUM nous permettent de comparer la confiance observable en sortie avec des indices fondés sur les activations internes.

1. Métriques mécanistiques

Parametric Force Score (PFS)

Le Parametric Force Score (PFS) quantifie l'amplitude de la mise à jour introduite par le Feed-Forward Network (FFN) dans le flux résiduel pour une couche et un token donnés :

$$PFS(t,l) = \left\| x^{(l,t)}_{\text{post-ffn}} - x^{(l,t)}_{\text{pre-ffn}} \right\|_2 .$$

Ici, $x^{(l,t)}_{\text{pre-ffn}}$ est la représentation du token avant le bloc FFN, $x^{(l,t)}_{\text{post-ffn}}$ est la représentation après le bloc FFN, l désigne la couche et t le token généré.

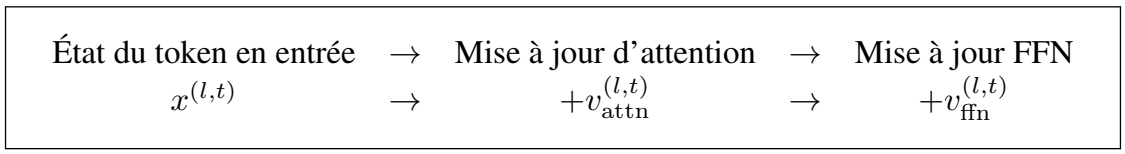

FIGURE 1 – Pipeline de traitement par attention et réseau feed-forward.

Pathway Alignment Score (PAS)

Le Pathway Alignment Score (PAS) mesure l'alignement entre la mise à jour d'attention et la mise à jour FFN :

$$PAS(t,l) = \frac{v_{\text{attn}}^{(l,t)} \cdot v_{\text{ffn}}^{(l,t)}}{\left\|v_{\text{attn}}^{(l,t)}\right\|_2 \left\|v_{\text{ffn}}^{(l,t)}\right\|_2}.$$

Les deux vecteurs de mise à jour sont définis comme suit :

$$v_{\text{attn}}^{(l,t)} = x_{\text{pre-ffn}}^{(l,t)} - x_{\text{input}}^{(l,t)}, \qquad v_{\text{ffn}}^{(l,t)} = x_{\text{post-ffn}}^{(l,t)} - x_{\text{pre-ffn}}^{(l,t)}.$$

Représentation matricielle de PFS et PAS

Pour chaque réponse générée, PFS et PAS sont stockés sous forme de matrices :

$$PFS \in \mathbb{R}^{T\times N}, \qquad PAS \in \mathbb{R}^{T\times N},$$

où T est le nombre de tokens générés et N le nombre de couches du transformer. Les lignes correspondent aux tokens et les colonnes aux couches. Comme ces matrices ne peuvent pas être utilisées directement comme entrées scalaires d'un classifieur, elles sont agrégées en scores de taille fixe. Nous avons testé plusieurs stratégies d'agrégation, notamment la moyenne globale, la médiane, la moyenne de la dernière couche, la moyenne top-k et la régression logistique couche par couche. **La moyenne de la dernière couche a produit les meilleurs résultats de classification fondée sur un seuil** ; cette agrégation est donc utilisée pour le score FACTUM (PFS+PAS) rapporté.

2. Métriques de référence fondées sur la confiance

Perplexity

La perplexité[7] mesure la vraisemblance d'une séquence générée selon le modèle :

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

Entropy

L'entropie[19] mesure l'incertitude dans la distribution du prochain token du modèle :

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

Energy Score

L'energy score[13] est calculé directement à partir des logits :

$$E_t = -\log\sum_{i=1}^{V}\exp(z_i^{(t)}).$$

$$\text{Energy} = \frac{1}{T}\sum_{t=1}^{T} E_t.$$

P(True) Score

$P(\text{True})$[9] estime la probabilité auto-évaluée par le modèle qu'une réponse générée soit correcte :

$$P(\text{True}) = p(\text{“true”} \mid \text{question}, \text{answer}, \text{prompt}).$$

3. Protocole d'évaluation

Orientation des scores

Chaque réponse générée reçoit une étiquette binaire, où 1 désigne une réponse correcte et 0 une réponse incorrecte. Comme les métriques évaluées ont des orientations différentes, les scores sont orientés avant l'application d'un seuil. Pour la perplexité, l'entropie et l'energy score, des valeurs plus faibles indiquent une confiance plus forte du modèle ; le score est donc inversé. Pour $P(\text{True})$, des valeurs plus élevées indiquent une confiance plus forte dans la correction de la réponse. Après orientation, toutes les métriques suivent la même convention : des valeurs plus grandes indiquent une preuve plus forte en faveur de la classe positive.

Formellement, pour un score s_i, le score orienté est :

$$\tilde{s}_i = \begin{cases} s_i & \text{si des valeurs plus élevées indiquent la correction,} \\ -s_i & \text{si des valeurs plus faibles indiquent la correction.} \end{cases}$$

Classification fondée sur un seuil

Pour chaque métrique, le score orienté $\tilde{s}_i$ est converti en prédiction binaire au moyen d'un seuil τ sélectionné sur le fold d'entraînement. Le seuil est choisi selon le critère J de Youden. Pour chaque seuil candidat θ_k, nous calculons le taux de vrais positifs TPR_k et le taux de faux positifs FPR_k, puis sélectionnons :

$$k^* = \arg\max_k \left(TPR_k - FPR_k\right), \qquad \tau^* = \theta_{k^*}.$$

Les prédictions sont ensuite obtenues ainsi :

$$\hat{y}_i = \begin{cases} 1 & \text{if } \tilde{s}_i \geq \tau^*, \\ 0 & \text{sinon.} \end{cases}$$

Ces prédictions sont comparées aux étiquettes de référence afin de calculer TP, TN, FP et FN, à partir desquels sont dérivés l'exactitude, la précision, le rappel et le score F1.

3 Jeu de données et contexte d'évaluation

Cette expérimentation s'appuie sur le benchmark TREC-COVID, une collection standard de recherche d'information conçue pour soutenir les travaux sur la récupération d'information et la réponse à des questions dans la littérature scientifique sur la COVID-19. Le benchmark est composé de trois éléments principaux :

1. **Topics et besoins informationnels.** TREC-COVID fournit un ensemble de topics, où chaque topic correspond à un besoin informationnel lié à la COVID-19. Chaque topic contient une requête et une brève description contextuelle précisant la portée visée par la question. Dans ce travail, nous avons considéré l'ensemble complet des 50 topics TREC-COVID, tels qu'ils sont fournis dans les fichiers officiels [17,16].
2. **Collection de documents.** La collection de documents est dérivée du corpus CORD-19, qui contient des articles scientifiques liés à la COVID-19 et aux coronavirus. Ce corpus a été constitué pour soutenir les recherches en fouille de textes et en recherche d'information sur la littérature COVID-19 [21,2]. Pour cette expérimentation, nous avons utilisé la version finale des métadonnées CORD-19 datée du 2022-06-02 [2].
3. **Jugements de pertinence.** TREC-COVID fournit des jugements de pertinence, communément appelés qrels [15]. Chaque entrée qrels contient quatre champs :
   — l'identifiant du topic ;
   — le champ d'itération, qui est un placeholder historique de TREC et est généralement ignoré par les outils d'évaluation ;
   — l'identifiant du document ;
   — le jugement de pertinence.

   Les étiquettes de pertinence sont définies comme suit :
   — 2 : document très pertinent ;
   — 1 : document partiellement pertinent ;
   — 0 : document non pertinent.

   Dans ce travail, nous avons utilisé le fichier qrels final du Round 5, daté du 2020-07-16, qui contient les jugements de pertinence pour l'ensemble des 50 topics TREC-COVID.

3.1 Pourquoi TREC-COVID seul était insuffisant

Bien que TREC-COVID constitue un benchmark précieux pour la recherche d'information biomédicale, il n'était pas suffisant à lui seul pour l'objectif de cette étude. Le benchmark a été conçu autour de topics de recherche d'information sur la COVID-19 et de jugements officiels de pertinence, notamment au fil de ses rounds successifs et des qrels du Round 5 [17,16,15]. Ces qrels fournissent des jugements de pertinence au niveau du topic pour des documents entiers, tandis que notre évaluation requiert des preuves au niveau du passage pouvant directement soutenir ou réfuter une réponse générée. Autrement dit, nous devions savoir non seulement si un article était globalement lié à un topic, mais aussi si un passage précis justifiait une réponse factuelle.

Cette distinction est importante, car un document peut être pertinent pour un topic sans répondre explicitement à la question. Inversement, un passage utile pour soutenir une réponse peut apparaître dans un document qui n'a pas été jugé dans les qrels officiels. S'appuyer uniquement sur les annotations originales de TREC-COVID rendrait donc difficile la comparaison des réponses générées par les modèles avec une vérité de référence claire.

Pour répondre à cette limite, nous avons construit manuellement un ensemble de documents et de passages soutenant les réponses pour les 50 questions TREC-COVID. Pour chaque topic, la question originale ou une version reformulée a été soumise à une interface de récupération fondée sur la similarité, construite sur la collection CORD-19 [21,2,6]. Les articles récupérés ont été classés au moyen de calculs de similarité sur les résumés, puis inspectés manuellement afin d'identifier l'article et le passage précis soutenant la réponse.

Lorsque les résultats de récupération CORD-19 ne fournissaient pas de preuves suffisamment explicites, nous avons recherché des sources autoritatives supplémentaires. Cela a conduit à l'inclusion de

quatre documents externes :

— *Clinical management of COVID-19 : living guideline, June 2025*, publié par l'Organisation mondiale de la santé (OMS).
— *Low-cost dexamethasone reduces death by up to one third in hospitalised patients with severe respiratory complications of COVID-19*, publié par l'Université d'Oxford en lien avec l'essai RECOVERY.
— *Impact of Coronavirus Disease 2019-Related School Closures on the Drivers of Child Health*, un article pédiatrique indexé dans PubMed Central.
— *mRNA Vaccines : Current Applications and Future Directions*, publié dans *MedComm* par Wiley et indexé dans PubMed.

Bien que seuls quatre documents externes aient été ajoutés, différentes sections de ces sources ont été utilisées pour soutenir neuf instances de réponse parmi les topics TREC-COVID. Ces ajouts n'avaient pas pour objectif de remplacer le benchmark, mais de l'adapter à un objectif d'évaluation différent : la vérification factuelle des réponses et l'analyse des hallucinations.

3.2 Comparaison avec les jugements de pertinence TREC-COVID

Pour évaluer la relation entre les preuves identifiées manuellement et les jugements officiels de TREC-COVID, nous avons associé les articles sélectionnés aux identifiants de documents CORD-19 lorsque cela était possible, puis nous les avons comparés aux qrels du Round 5 [16,15].

Le jeu de données contient 50 questions avec réponses. Pour ces questions, notre processus manuel a identifié 168 articles soutenant les réponses. Parmi eux, 159 articles ont pu être associés à des documents de la base TREC-COVID/CORD-19 [21,2], tandis que 9 articles n'ont pas été retrouvés dans la collection du benchmark.

Après suppression des paires requête-document dupliquées, la comparaison a produit 149 paires requête-document uniques. Leur distribution par rapport aux qrels officiels de TREC-COVID était la suivante :

— 5 documents ont été jugés très pertinents ;
— 3 documents ont été jugés partiellement pertinents ;
— 7 documents ont été jugés non pertinents ;
— 134 documents n'étaient pas jugés.

Le grand nombre de documents non jugés est important. Ces documents étaient présents dans CORD-19 et ont été vérifiés manuellement comme contenant des passages soutenant les réponses, mais ils n'ont pas reçu d'étiquettes de pertinence dans les qrels officiels du Round 5 [15]. Cela ne signifie pas qu'ils sont non pertinents ; cela indique plutôt qu'ils se trouvaient en dehors de l'ensemble évalué pour les topics correspondants.

Cette comparaison suggère que notre ensemble de preuves curé complète les qrels originaux de TREC-COVID en identifiant des passages précis soutenant les réponses, y compris dans des documents qui n'ont pas été jugés par le benchmark. Toutefois, les documents non jugés ne doivent pas être considérés comme automatiquement pertinents. Leur utilité dans notre jeu de données provient de la vérification manuelle au niveau du passage effectuée dans ce travail.

Topic	Question	CORD doc ID	Passage Identifié	Discussion
1	what is the origin of COVID-19	dc30gkfe	*Coronavirus disease 2019 (COVID-19) is a current new virulent disease rising its transmission and fatality with each passing day in the worldwide population. COVID-19 is emerged as a respiratory infection and a suspicious origin of animals and transmission to human in Wuhan, China on December 2019. Later this, the virus was transmitted from person to person through droplets and contacts.*	The passage is relevant because it directly addresses the origin and early transmission of COVID-19. It mentions Wuhan, December 2019, a suspected animal origin, and subsequent human-to-human transmission. Although the document was judged irrelevant in the official qrels, this passage provides usable evidence for an answer about the origin of COVID-19.
3	will SARS-CoV2 infected people develop immunity ? Is cross protection possible ?	1waejlwb	*The hypothesis that BCG may offer protection from COVID-19. Heterologous protection offered by BCG through production of trained immunity, epigenetic reprogramming of monocytes, non-specific activation of NK cells, and increase of pro-inflammatory cytokines [...] may be the mechanism behind its cross-protection against the novel coronavirus.*	The passage is relevant to the cross-protection component of the question. It discusses the possibility that the Bacillus Calmette–Guérin vaccine may provide non-specific immune protection against COVID-19 through trained immunity and immune-cell activation. It does not fully answer whether infected individuals develop immunity, but it supports the part of the query concerning cross-protection.
16	how long does coronavirus remain stable on surfaces ?	0y8lfjkx	*The time required for the virus titer to decrease 99.9% shows that in tap water, coronaviruses are inactivated faster in water at 23°C, 10 days, than in water at 4°C, more than 100 days. Coronaviruses die off rapidly in wastewater, with values between 2 and 4 days.*	The passage is partially relevant because it concerns coronavirus environmental stability and survival time. However, the evidence refers to water and wastewater rather than physical surfaces. It is therefore useful as background evidence on coronavirus persistence, but it does not directly answer the surface-stability aspect of the topic.
17	are there any clinical trials available for the coronavirus	804r5xzd	*Lopinavir/Ritonavir, nucleoside analogues, neuraminidase inhibitors, Remdesivir, peptide EK1, abidol, RNA synthesis inhibitors, anti-inflammatory drugs, and Chinese traditional medicine could be drug treatment options for 2019-nCoV. However, the efficacy and safety of these drugs still need to be further confirmed by clinical experiments.*	The passage is partially relevant because it identifies candidate treatments for coronavirus and explicitly states that their efficacy and safety require confirmation through clinical experiments. However, it does not list active or completed clinical trials. It therefore supports the broader clinical-trial motivation of the query, but only indirectly.
19	what type of hand sanitizer is needed to destroy Covid-19 ?	fjiffj86	*In order to prevent spread of COVID-19, World Health Organization has specified that measures such as cleaning hands regularly with alcohol-based hand sanitizer or washing with soap and water, avoiding touching nose, eyes, mouth and social distancing should be followed.*	The passage is partially relevant because it explicitly mentions alcohol-based hand sanitizer as a preventive measure recommended by the WHO. However, it does not specify ethanol or isopropanol concentration, nor does it provide experimental evidence about viral inactivation. It therefore supports the answer at a general level but not at a precise formulation level.
19	what type of hand sanitizer is needed to destroy Covid-19 ?	ji5pqh47	*Hand hygiene is of utmost importance for the prevention of COVID-19 among health care workers. This purpose can be achieved by applying alcohol-based hand rubs, washing hands properly with soap and water, and applying other antiseptic agents.*	This passage is also partially relevant. It supports the use of alcohol-based hand rubs and antiseptic agents for COVID-19 prevention, but it does not specify the exact chemical composition or concentration required to destroy SARS-CoV-2. As with the previous case, the passage is useful but incomplete for the precise question.
28	what evidence is there for the value of hydroxychloroquine in treating Covid-19 ?	0eizsamh	*There is currently no robust evidence to support prescribing hydroxychloroquine as a treatment or prophylaxis for COVID-19.*	The passage is highly relevant because it directly addresses the evidence for hydroxychloroquine in COVID-19 treatment and prevention. It states that robust evidence is lacking, which provides a clear answer to the question. Although the document was judged irrelevant in the official qrels, this passage is directly useful for factual answer verification.

TABLE 1 – Passages manuellement pertinents dans des documents TREC non pertinents

4 Méthodologie

Notre méthode simule un cadre dans lequel un modèle peut avoir été exposé à du contenu scientifique protégé par des restrictions de propriété intellectuelle. Pour construire ce cadre, nous utilisons un jeu de données COVID-19 dérivé de PubMed et de sources biomédicales connexes, où la réutilisation automatisée du contenu des articles peut être restreinte par les licences des éditeurs ou des plateformes [18]. À partir de ce matériau, nous avons construit un benchmark personnalisé composé de questions sur la COVID-19, de réponses de référence et de passages justificatifs vérifiés manuellement.

Nous avons évalué trois modèles à poids ouverts : LLaMA 3.1 8B, LLaMA 3.2 3B et Ministral 3-3B. Pour chaque modèle, nous avons généré des réponses aux questions du benchmark et attribué des étiquettes binaires indiquant si la réponse était correcte ou incorrecte au regard des preuves vérifiées. Pendant l'inférence, nous avons calculé à la fois des baselines fondées sur la confiance et des métriques internes de type FACTUM, en nous concentrant sur PFS et PAS.

Tous les scores de confiance ont été calculés avec le même modèle que celui ayant généré la réponse. Cela signifie que chaque modèle a été évalué comme son propre juge plutôt qu'au moyen d'un modèle externe. Ce choix est important, car la perplexité, l'entropie, l'énergie, $P(\mathrm{True})$, PFS et PAS visent à refléter la confiance et le comportement interne du modèle générateur lui-même. Utiliser un évaluateur tiers mesurerait plutôt l'incertitude du modèle juge lors de la lecture de la réponse, et non la confiance du modèle qui l'a produite.

5 Résultats

Les valeurs de PFS et PAS ont été extraites pendant la génération et stockées sous forme de matrices token-par-couche. Comme ces matrices doivent être réduites à des caractéristiques scalaires pour la classification, nous avons comparé plusieurs stratégies d'agrégation : moyenne globale, médiane, moyenne de la dernière couche, moyenne top-k et régression logistique couche par couche. La moyenne de la dernière couche a donné les meilleurs résultats fondés sur un seuil et a donc été utilisée pour le score FACTUM (PFS+PAS) rapporté.

Les résultats montrent que FACTUM (PFS+PAS) obtient des performances compétitives par rapport aux baselines classiques de confiance, et parfois supérieures. Cela suggère que les signaux fondés sur les activations internes fournissent des informations qui ne sont pas entièrement capturées par les métriques au niveau de la sortie, telles que la perplexité, l'entropie, l'énergie ou $P(\mathrm{True})$. En particulier, la meilleure performance de l'agrégation sur la dernière couche indique que le signal le plus utile s'exprime près de la décision de sortie du modèle.

Métrique / Baseline	AUC	Exactitude	Précision	Rappel	F1
Perplexity	0.4448	0.43	0.3146	0.7117	0.4314
LN-Entropy	0.5948	0.554	0.4058	0.7883	0.5260
Energy Score	0.6563	0.586	0.4295	0.7750	0.5399
P(True) Query	0.6395	0.572	0.3661	0.4533	0.3468
FACTUM (PFS+PAS)	0.6732	0.6954	0.6400	0.6500	0.5927

TABLE 2 – Résultats on LLaMA 3.1 8B.

Métrique / Baseline	AUC	Exactitude	Précision	Rappel	F1
Perplexity	0.5433	0.532	0.3594	0.6200	0.4360
LN-Entropy	0.5095	0.468	0.2619	0.4267	0.3019
Energy Score	0.4143	0.442	0.2495	0.4867	0.3224
P(True) Query	0.5343	0.450	0.2540	0.5667	0.3440
FACTUM (PFS+PAS)	0.5952	0.5840	0.4070	0.6600	0.4773

TABLE 3 – Résultats on LLaMA 3.2 3B.

Métrique / Baseline	AUC	Exactitude	Précision	Rappel	F1
Perplexity	0.5600	0.524	0.2634	0.6900	0.3629
LN-Entropy	0.4975	0.430	0.2013	0.7200	0.3116
Energy Score	0.4600	0.474	0.1305	0.3800	0.1895
P(True) Query	0.4312	0.572	0.1521	0.3300	0.1993
FACTUM (PFS+PAS)	0.5788	0.4740	0.2330	0.6400	0.3266

TABLE 4 – Résultats on Ministral 3B.

6 Discussion

La collection TREC-COVID est principalement organisée autour d'articles scientifiques et de leurs métadonnées associées, y compris les résumés. Dans notre workflow, ces résumés ont été utiles pour construire un processus de recherche fondé sur la similarité, permettant d'identifier des documents candidats pour répondre aux questions TREC-COVID et évaluer les réponses générées par les modèles. Cependant, les résumés étaient souvent insuffisants pour déterminer si une réponse était factuellement correcte ou correctement étayée. Cela a nécessité une inspection plus approfondie du contenu intégral des articles, car les preuves pertinentes apparaissaient fréquemment dans le corps principal de l'article plutôt que dans le résumé.

Pour cette raison, le processus de vérification manuelle a fréquemment nécessité des recherches web supplémentaires utilisant des éléments du topic, de la réponse du modèle et de concepts scientifiques candidats. Cela a parfois conduit à des articles externes soutenant la réponse, même lorsque la preuve n'était pas directement disponible dans les métadonnées de résumés CORD-19. Comme discuté précédemment, ces sources externes n'ont été utilisées que lorsqu'elles fournissaient des passages explicites soutenant la réponse. Toutefois, cela soulève aussi une question plus large sur l'origine de certaines réponses de modèles. Plusieurs éditeurs et revues scientifiques restreignent l'accès automatisé à leur contenu intégral, y compris le scraping pour l'entraînement de modèles à grande échelle. Néanmoins, certaines réponses générées montrent un fort chevauchement lexical et structurel avec des contenus scientifiques protégés ou à accès restreint. Cela ne prouve pas en soi une exposition aux données d'entraînement, mais motive une investigation plus approfondie des comportements de type mémorisation.

Un cas illustratif concerne la réponse générée pour le Topic 38, qui porte sur la réponse inflammatoire et la pathogenèse de la COVID-19. Nous utilisons cet exemple comme cas qualitatif de similarité entre une réponse générée et un article externe rédigé par des humains sur le traitement par azvudine[11]. La comparaison ci-dessous met en évidence plusieurs segments étroitement alignés.

Segment de réponse du modèle	Segment de source rédigé par un humain
“the virus enters the host cell through the interaction between its spike protein and the angiotensin-converting enzyme 2 (ACE2) receptor”	“the spike protein (S) of SARS-CoV-2 interacts with the cellular receptor ACE2”
“the viral genome is released from the viral capsid and is translated into viral proteins by the host cell machinery”	“The viral genome is released in the cytoplasm and translated into the viral replicase polyproteins”
“released from the host cell through the process of exocytosis”	“released from the host’s infected cells into the extracellular space by exocytosis”

TABLE 5 – Segments alignés entre la réponse du modèle pour le Topic 38 et un passage d’une source externe.

Cette comparaison doit être interprétée avec prudence. Le contenu aligné correspond à des connaissances scientifiques standards et peut apparaître dans plusieurs sources biomédicales. Par conséquent, la similarité observée ne suffit pas à elle seule à établir une mémorisation. Toutefois, d’un point de vue computationnel, l’exemple est notable parce que la réponse générée préserve à la fois les ancrages lexicaux et l’organisation séquentielle du passage externe. Cela motive une analyse plus poussée sur un ensemble plus large d’exemples et avec des mesures de similarité plus systématiques, afin de distinguer les connaissances ordinaires du domaine, le chevauchement fortuit et la mémorisation possible de contenus protégés ou à accès restreint.

Références

[1] AGRAWAL A., SUZGUN M., MACKEY L. & KALAI A. (2024). Do language models know when they’re hallucinating references ? In *Findings of the Association for Computational Linguistics : EACL 2024*, p. 912–928.

[2] ALLEN INSTITUTE FOR AI (2022). The covid-19 open research dataset (cord-19). https://github.com/allenai/cord19. Final release dated 2022-06-02.

[3] CHEN Y. *et al.* (2024). INSIDE : LLMs’ internal states retain the power of hallucination detection.

[4] 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*.

[5] 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*.

[6] ERIC SANJUAN - LIA AVIGNON (2026). Trec covid-19 demo. https://dev.ai-edge.eu/dataviz/covid19/.

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

[8] JI Z., CHEN D., ISHII E., CAHYAWIJAYA S., BANG Y., WILIE B. & FUNG P. (2024). Llm internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop*, p. 88–104.

[9] 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*.

[10] KOVÁCS Á. & RECSKI G. (2025). Lettucedetect : A hallucination detection framework for rag applications. *arXiv preprint arXiv :2502.17125*.

[11] LI J., ZHU B., LU J., DONG Z., LI P., LI W., ZHENG C., CHANG J. & SHANG S. (2025). Advances in the effectiveness and safety of azvudine treatment : a comprehensive review. *Frontiers in Pharmacology*, **16**. DOI : 10.3389/fphar.2025.1524072.

[12] LIN S., HILTON J. & EVANS O. (2022). TruthfulQA : Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, p. 3214–3252 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.229.

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

[14] MUSIC BUSINESS WORLDWIDE (2024). Gema wins landmark ruling against openai over chatgpt's use of song lyrics.

[15] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2020a). Trec-covid round 5 relevance judgments. https://ir.nist.gov/trec-covid/qrels5.html.

[16] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2020b). Trec-covid round 5 task guidelines. https://ir.nist.gov/trec-covid/round5.html.

[17] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2020c). Trec-covid topics. https://ir.nist.gov/trec-covid/data/topics-rnd1.xml.

[18] PUBMED CENTRAL (2026). Copyright information for pubmed central (pmc).

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

[20] TONEY-WAILS A. & WAILS R. (2025). Certain but not probable ? differentiating certainty from probability in llm token outputs for probabilistic scenarios.

[21] WANG L. L., LO K., CHANDRASEKHAR Y., REAS R. *et al.* (2020). CORD-19 : The covid-19 open research dataset. *arXiv preprint arXiv :2004.10706*.