Aligner méthode historique et RAG : transformer un assistant conversationnel en chaîne de preuve auditable et discutable

Marie Puren[1,2], Donghan Bian[2,1], Aurélien Pellet[1], Julien Perez[1], and Florian Cafiero[1]

[1] Laboratoire de recherche d'Epita (LRE), EPITA, Le Kremlin-Bicêtre, France
[2] Centre Jean-Mabillon, Ecole nationale des chartes, Paris, France

Abstract

This article examines the challenges raised by deploying Retrieval-Augmented Generation (RAG) systems for the exploration of digitized historical sources. Starting from the observation that disciplinary acceptance remains fragile, it asks the following question: how can a RAG system applied to noisy and heterogeneous archives ensure conditions of verification and critique that are compatible with the historical method? Presented as a position paper, this paper offers a conceptual framing and preliminary directions for guiding the development of RAG devices aligned with these requirements. It argues for restoring control over the interpretive chain by articulating three conditions: traceability (the ability to locate documents and passages precisely), auditability (the ability to inspect the transformations and parameters throughout the pipeline), and discussability (the ability to subject statements to debate by distinguishing evidence from interpretation). Its main contribution is an auditability framework that translates historians' requirements into instrumented conditions: (1) documentary anchoring (provenance and integrity), (2) explicit separation of quotation, paraphrase, and inference, (3) restoration of context and plurality of sources, (4) traceability of execution conditions and error diagnosis (retrieval vs. generation), and (5) abstention mechanisms when the evidence is insufficient.

Mots-clés: histoire, génération augmentée par récupération
Keywords: history, retrieval augmented generation

1 RAG et histoire : tensions méthodologiques et enjeux de confiance

Les systèmes de Génération Augmentée par Récupération (*Retrieval Augmented Generation* ou RAG) combinent recherche dans un corpus et génération de réponses à partir de passages récupérés (Figure 1) [20]. De plus en plus mobilisés pour explorer des corpus de sources historiques numérisées (cf. section 2.1), ces dispositifs permettent d'interroger rapidement de vastes ensembles hétérogènes en langage naturel et de produire des synthèses appuyées sur des extraits cités, utiles pour s'orienter dans les archives et dégager des pistes d'analyse ; ils constituent ainsi un nouvel outil à la disposition des historiens pour « lire à distance » leurs corpus [24].

L'acceptabilité de ces outils reste toutefois fragile dans les communautés historiennes : une erreur visible peut suffire à discréditer l'ensemble du dispositif, faute d'un cadre partagé permettant de distinguer une erreur ponctuelle, une zone d'incertitude où plusieurs lectures sont possibles, et un cas hors périmètre où le système ne dispose pas des éléments nécessaires pour répondre de manière fondée. Cette fragilité renvoie aussi à des tensions avec des normes professionnelles centrales en histoire [31] : la critique des sources, l'attention à la matérialité des documents et de leurs

Marie Puren, Donghan Bian, Aurélien Pellet, Julien Perez, and Florian Cafiero. "Aligner méthode historique et RAG : transformer un assistant conversationnel en chaîne de preuve auditable et discutable." *Actes de la Conférence Humanistica*, éd. par Serena Crespi, Simon Gabay, Martin Grandjean, Ariane Pinche, Marie Puren et Léa Saint-Raymond. Vol. 4. Anthology of Computers et the Humanities. 2026, 1–9. https://doi.org/10.63744/G14mF7UizFzS.

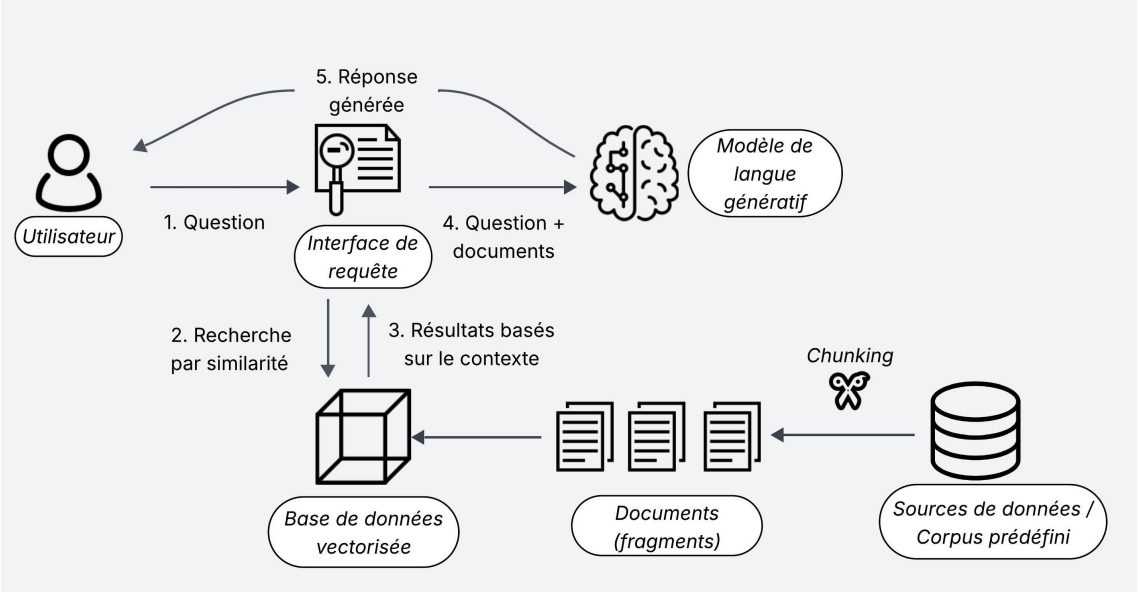

FIGURE 1 – Architecture d'un système RAG

contextes de production, et la valeur accordée à l'enquête historique. Or le RAG tend à déplacer ces gestes vers une chaîne computationnelle dont les opérations ne sont pas immédiatement visibles. Parce qu'ils dialoguent de façon fluide, ces assistants peuvent aussi produire une illusion de solidité : une réponse bien formulée ressemble à un savoir établi, même quand elle repose sur des indices fragiles [34]. Si l'interface n'affiche pas clairement les limites, les conditions de validité et la provenance des éléments mobilisés, l'usage tend à osciller entre une acceptation pragmatique et un rejet brutal dès qu'une erreur est repérée – d'où l'exigence de rendre ses médiations examinables [17].

L'enjeu n'est pas seulement que le RAG se trompe parfois, mais que la preuve est médiée par une chaîne computationnelle – de la transcription à la segmentation, en passant par l'indexation, la récupération et la reformulation – difficile à inspecter pas à pas. Autrement dit, la critique ne porte plus seulement sur le document-source, mais aussi sur les opérations qui rendent ce document lisible et mobilisable par le système, et sur leurs effets. La question qui nous guide est donc la suivante : **comment garantir, avec un RAG appliqué à des sources anciennes numérisées, des conditions de vérification et de critique compatibles avec la méthode historique ?** L'objectif est de reconstruire un contrôle épistémique : pouvoir comprendre sur quoi repose un énoncé, le vérifier et le contester. Pour cela, nous distinguons la traçabilité (retrouver précisément les documents et passages mobilisés), l'auditabilité (rendre inspectables les transformations et paramètres qui ont conduit à ces passages et à leur reformulation), et la discutabilité (mettre l'énoncé en débat sur la base de ces pièces, en séparant preuve et interprétation).

Dans la suite de cet article, nous adoptons une démarche d'exposé de position : ce texte propose un cadrage méthodologique et des pistes préliminaires pour orienter le développement de dispositifs RAG alignés avec les exigences de la méthode historique. Nous soutenons que la confiance se construit à condition de faire du RAG une chaîne de preuve. Autrement dit, chaque énoncé doit être relié à des passages identifiables, replacés dans leur contexte, et assortis d'indices sur leur sélection et leurs transformations. L'objectif est d'en faire un instrument de travail historien, au-delà d'un outil d'accès rapide aux archives : le RAG doit donc être contrôlable, critiquable, reproductible et appropriable. Nous proposons en conséquence un cadrage et un protocole minimal d'auditabilité alignés sur les exigences de la méthode historique [5 ; 18 ; 32].

2 Interroger les archives avec le RAG : état des lieux

2.1 RAG pour les corpus patrimoniaux

Les usages du RAG se multiplient pour interroger des corpus patrimoniaux, souvent via des assistants conversationnels [10 ; 15 ; 16 ; 19 ; 21 ; 23 ; 37 ; 38]. Or, ces systèmes sont principalement conçus pour des corpus contemporains et relativement propres, tandis que les archives historiques sont hétérogènes, parfois multilingues, longues, et fréquemment bruitées. Une partie de la littérature récente souligne que, dès que l'on se confronte à des usages patrimoniaux et multimodaux, la fiabilité et l'auditabilité deviennent pourtant des objectifs explicites [1]. Plus généralement, les méthodes computationnelles transforment les pratiques et les objets ; elles exigent donc des critères de légitimation disciplinaires, et pas seulement des scores [2]. Enfin, l'explosion des recherches sur le RAG ne s'accompagne pas encore d'une standardisation stable des régimes d'évaluation et d'auditabilité, ce qui plaide pour des protocoles situés, répondant aux exigences du domaine, plutôt que pour un score unique [6].

2.2 Récupération vs génération : diagnostiquer les erreurs du RAG

Nous pouvons distinguer deux dimensions à évaluer : (1) la récupération (*retrieval*), c'est-à-dire la capacité du système à retrouver des documents et fragments pertinents ; (2) la génération, c'est-à-dire la manière dont le système formule une réponse à partir de ces fragments. Cette distinction est utile si elle permet un diagnostic : un problème peut venir d'une récupération inadéquate (mauvais passages) ou d'un glissement interprétatif lors de la génération (surinterprétation, généralisation, confusion). En pratique, plusieurs frictions techniques ont un effet directement épistémique : la segmentation (*chunking*) peut perdre le contexte ; le bruit OCR/HTR peut altérer la lecture et déplacer la critique vers la chaîne de transcription ; la sensibilité au prompt et à la configuration fragilise la reproductibilité ; enfin, le mélange de citation, paraphrase et interprétation brouille la frontière entre preuve et commentaire.

Les juges LLM (*LLM-as-a-Judge*) [41] peuvent aider à classer des erreurs ou à comparer des réponses, à condition d'être calibrés et intégrés comme auxiliaires de critique, non comme arbitres de vérité [8]. Les résultats d'évaluation montrent toutefois une forte sensibilité aux tâches, aux domaines et aux hyperparamètres ; cela plaide pour des protocoles à visée diagnostique – cartographier les conditions de succès/échec et les profils d'erreurs – plutôt que pour des classements globaux peu transférables [9]. Enfin, les critiques des « *context metrics* » rappellent que des dimensions cruciales, comme la disciplinarité, l'interprétation et le contexte, résistent à une quantification simple : les métriques doivent rester arrimées à des pratiques de lecture et de validation [3]. Si la réduction des hallucinations repose sur de nombreuses solutions techniques (contraindre la récupération, organiser l'abstention lorsque les preuves sont insuffisantes, ou encore ajouter des étapes de vérification) [40], l'objectif central en histoire n'est pas d'optimiser une métrique globale, mais de rendre la preuve discutable : expliciter les appuis documentaires, signaler les zones d'incertitude, et borner clairement ce qui peut, ou ne peut pas, être inféré. Ce besoin de contrôle sur la chaîne RAG ne concerne pas seulement l'histoire, mais plus largement tout domaine de connaissance où la réponse doit pouvoir être justifiée, vérifiée et contestée à partir de sources explicites [14 ; 25 ; 33].

3 Une grille d'auditabilité pour le RAG en histoire

L'enjeu n'est pas seulement l'optimisation technique de la chaîne, mais l'explicitation, étape par étape, des transformations susceptibles d'affecter le statut de preuve des éléments mobilisés. La transparence attendue doit être opératoire : exposer les décisions prises pour la segmentation, l'indexation, les paramètres de récupération et les contraintes de génération, ainsi que leurs effets,

doit permettre à l'utilisateur d'agir – en reformulant ou comparant – plutôt que de consommer un résultat présenté comme définitif [33].

Notre contribution consiste à traduire les exigences historiennes en conditions instrumentées, afin de produire des sorties qui soutiennent une lecture critique, traçable, discutée et contextualisée [12 ; 26]. Nous proposons, dans cette perspective, une grille d'auditabilité qui précise ce qui doit être visible, traçable et vérifiable afin que la réponse puisse être considérée comme une preuve discutable. Cette grille s'appuie sur des observations et des réflexions « de terrain », issues d'expérimentations menées sur des débats parlementaires et de la presse de la Troisième République (1870-1940) [4 ; 27 ; 28 ; 29], deux corpus longs, bruités et fortement dépendants du contexte, où la segmentation, le bruit OCR et la multiplicité des versions exacerbent les enjeux de surinterprétation, de traçabilité et de reproductibilité.

3.1 Ancrer la source : provenance, intégrité et traçabilité documentaire

La critique externe des sources exige des identifiants stables, des métadonnées explicites (document, édition, date, cote/identifiant, version), et la transparence des transformations (OCR/HTR, normalisation, nettoyage, découpage). Elle suppose aussi une documentation manipulable du corpus, comprenant les choix de numérisation, les lacunes, les conditions de production, et impliquant les rôles et arbitrages des institutions documentaires [17]. Un risque spécifique de l'utilisation des grands modèles de langue est la production de références plausibles mais invérifiables ; d'où la nécessité d'ancrages stables comme des liens et des identifiants, et de mécanismes de prévention des références non vérifiables, afin que toute citation puisse être retrouvée et contrôlée [7]. Enfin, l'opacité de la sélection des sources empêche la critique ; rendre la provenance consultable et manipulable est une condition d'appropriation [34].

3.2 Séparer preuve et commentaire : citation, paraphrase et inférence

Afin d'éviter la confusion entre preuve et commentaire, il semble nécessaire de distinguer explicitement ce qui est cité, paraphrasé ou inféré, et vérifier la fidélité de la réponse par un alignement réponse-passages récupérés, avec détection de glissements et surinterprétations. La fidélité (*faithfulness*) ne doit pas seulement signifier que des sources sont affichées : elle doit signifier que la réponse dépend effectivement de ces passages, et qu'on peut voir précisément quelles parties de l'énoncé sont soutenues par quels extraits [39].

3.3 Restituer le contexte : voisinage des fragments et pluralité des sources

Dans un système de génération augmentée par récupération, les grands corpus qui constituent la base de l'analyse sont stockés sous forme de fragments, appelés aussi segments ou *chunks*, dont la taille dépend de la stratégie de pré-traitement retenue. Cette segmentation soulève toutefois plusieurs défis. En particulier, elle peut fausser l'interprétation en isolant des indices de leur voisinage contextuel. Une exigence minimale consiste donc à garantir l'accès au contexte entourant chaque fragment (avant/après, structure du document, voire document entier si nécessaire), tout en encourageant le croisement des sources : multiplicité des passages, des documents et des points de vue, y compris contradictoires, plutôt qu'une réponse unique, énoncée comme une vérité univoque. Des choix d'interface tels que les « réponses par source » ou la comparaison explicite des sources soutiennent ainsi des pratiques proches du croisement historien [33].

3.4 Reproduire l'analyse : traçabilité des conditions d'exécution et diagnostic d'erreurs

La discussion des énoncés suppose de pouvoir reconstituer leurs conditions de production : prompts, configurations, versions de modèles, paramètres et index. Cette traçabilité des conditions

d'exécution permet de rejouer l'analyse, de tester la sensibilité (top-k[1], taille des segments) et de distinguer un effet de corpus d'un effet de paramétrage. Sur cette base, l'analyse d'erreurs nécessite d'être structurée en distinguant récupération et génération, par type de source, période ou genre, et en précisant la place et la forme de la vérification humaine. Elle doit également intégrer des mécanismes d'abstention lorsque les passages récupérés n'apportent pas de preuve suffisante [25]. La qualité des données en entrée doit être traitée comme un axe explicite, car un OCR médiocre peut faire chuter les performances et expliquer des réponses erronées [4 ; 30].

4 Montrer les limites pour fonder la confiance

Au-delà des métriques et de la traçabilité, l'objectif est de faire de la sortie RAG un support de raisonnement historien. Nous entendons par là un ensemble de mécanismes qui apprennent à l'utilisateur à lire la sortie du système comme un raisonnement révisable en distinguant (1) ce qui est établi par citation, (2) ce qui est inféré et (3) ce qui ne peut pas être arbitré avec les éléments disponibles et qui exige un retour au document. Il s'agit ainsi de répondre à l'illusion de certitude conversationnelle par des interactions qui incitent à vérifier, affichent les conditions de validité et signalent les zones d'ombre [34], tout en favorisant l'exploration d'alternatives [33]. L'explicitation des limites et des incertitudes n'est pas un aveu d'échec : c'est ce qui rend l'énoncé discutable. Dans cette perspective, l'erreur algorithmique peut même être méthodologiquement productive si elle est rendue visible et analysable, car elle signale des zones de complexité interprétative [35]. La confiance dépend ainsi de la robustesse des étapes de la chaîne de traitement et de la discutabilité des preuves et non d'une rhétorique de performance [11] : elle relève d'un régime de travail (contrôle, critique, discussion), et non d'une croyance dans l'outil.

Dans cette perspective, on peut dégager plusieurs principes. Par exemple, une typologie simple pourrait structurer l'usage : un énoncé produit par le RAG pourrait être « attesté » (soutenu par des passages cités), « inféré » (déduit à partir d'indices avec raisonnement explicitable), « indécidable » (non attestable avec le corpus récupéré, le niveau de bruit ou le contexte). Cette dernière catégorie doit être associée à des mécanismes d'abstention lorsque l'évidence est faible, afin de prévenir des synthèses plausibles mais non prouvées [25]. Ensuite, du côté de l'interface, l'objectif est d'inciter à la vérification, avec une présentation centrée sur les sources qui affiche d'abord les passages et leurs métadonnées, puis la synthèse. Des marqueurs doivent signaler les zones à risque (OCR/HTR, segmentation, ambiguïtés contextuelles) et l'interface doit proposer des alternatives lorsque plusieurs lectures sont plausibles. Enfin, les visualisations d'incertitude doivent éviter l'équation implicite « score = vérité » et privilégier des signaux utiles à l'enquête (points à vérifier, contradictions, fragilités OCR) [13].

Un tel cadrage rend la sortie discutable et révisable, compatible avec la critique interne et externe. Il facilite une validation collective, en partageant non seulement un résultat mais aussi son statut et ses zones d'incertitude, ce qui relève d'une reproductibilité adaptée aux exigences des humanités numériques [36]. Il contribue également à une reconnaissance disciplinaire du travail numérique comme travail historien, en réinstallant des gestes de critique et de justification.

5 Aligner le RAG sur la logique de preuve : de la grille au protocole

Il ne s'agit donc plus d'optimiser uniquement une performance de questions-réponses, mais de rendre l'énoncé historique auditable et discutable. Dans un esprit d'exposé de position, nous avançons un cadrage destiné à être discuté et mis à l'épreuve : la confiance relève moins d'une adhésion à l'outil que d'un dispositif soutenant l'herméneutique critique et la discutabilité de la preuve, dans le prolongement des débats en humanités numériques sur les inférences computationnelles et leurs

1. Désigne le nombre k de passages (ou documents) que l'étape de récupération renvoie au modèle de génération.

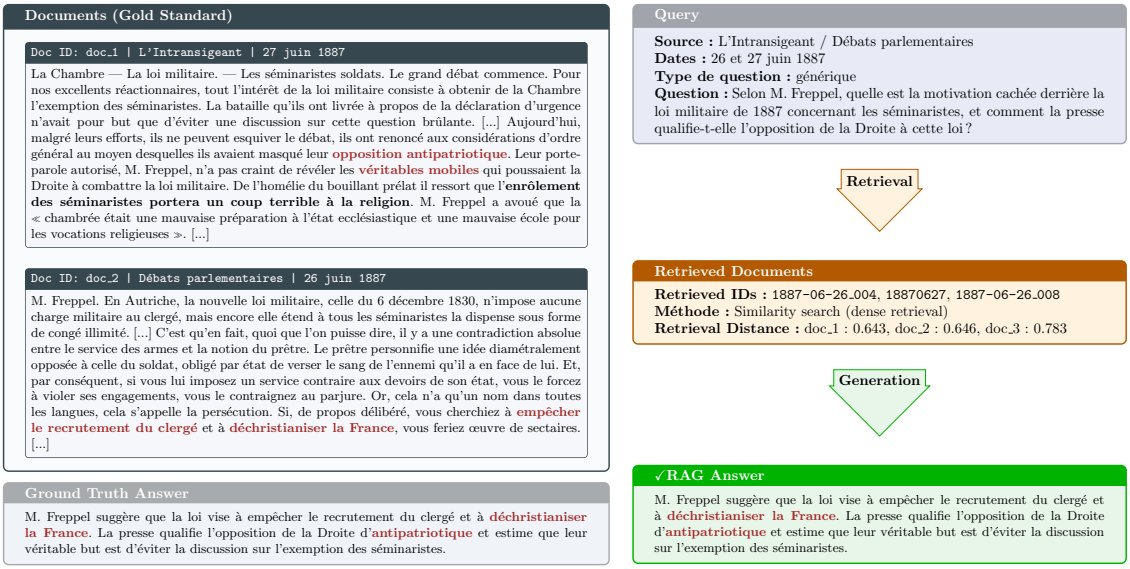

FIGURE 2 – Un extrait de question multi-hop question généré avec la chaîne de traitement RAG à partir de débats parlementaires et presse (1887) (cité dans [29])

effets sur les pratiques historiennes [11 ; 26 ; 31]. La proposition centrale est une grille d'auditabilité, lisible par des non-spécialistes, qui explicite ce qui doit être observable et traçable afin de soutenir une critique procédurale et l'appropriation des outils [17]. Elle implique aussi que les historiens demeurent acteurs de la chaîne de traitement, en refusant une externalisation qui dissocierait corpus, données et responsabilité scientifique. Enfin, nous esquissons une trajectoire de formalisation de cette grille en protocole minimal d'alignement méthode historique–RAG, et proposons des principes pour construire des jeux d'évaluation orientés vers la pratique historienne (comme l'illustre la Figure 2), afin d'arrimer l'évaluation à une logique de preuve plutôt qu'à des classements [28 ; 30].

Il ne s'agit pas de normaliser la démarche de l'enquête historienne, mais de fixer un socle commun : documenter les transformations des sources, les paramètres qui structurent récupération et génération, les limites connues, et les vérifications requises pour traiter une réponse comme un élément de preuve discutable. Cette trajectoire suppose une gouvernance partagée et une collaboration effective entre histoire et informatique, conçue comme co-définition des choix plutôt que comme prestation [22].

Références

[1] ABDELFATTAH, Ahmed M. H. et ATEF, David George. « From Recognition to Reliability : A Framework for Trustworthy Multimodal AI in Cultural Heritage (The MuseePal Case) ». In : *IADIS International Journal on WWW/Internet 23*, no. 2 (2025), p. 81-92. URL : `http://www.iadisportal.org/ijwi/papers/2025230206.pdf`.

[2] ARMAND, Cécile et HENRIOT, Christian. « Beyond Digital Humanities Thinking Computationally : A Position Paper ». In : *Beyond Digital Humanities : How Computational Methods Are Reshaping Scholarly Research*. Aix-en-Provence, France, 2023. URL : `https://shs.hal.science/halshs-04194570`.

[3] BARNETT, Tully et SUMNER, Tyne Daile. « Some Things Can't Be Measured : Rethinking Context, Metrics, and Disciplinarity in the Digital Humanities ». In : *The Compa-*

nion to Digital Humanities in Practice. Londres : Routledge, 2025. DOI : 10 . 4324 / 9781003327677-25.

[4] BIAN, Donghan, PUREN, Marie et CAFIERO, Florian. « How to Efficiently Explore Noisy Historical Data ? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG ». In : *Proceedings of the 10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 2026*, sous la dir. de Diego ALVES, Yuri BIZZONI, Stefania DEGAETANO-ORTLIEB, Anna KAZANTSEVA, Janis PAGEL et Stan SZPAKOWICZ. Rabat, Maroc : Association for Computational Linguistics, 2026, p. 241-250. DOI : 10 . 18653/v1/2026.latechclfl-1.23.

[5] BLOCH, Marc. *Apologie pour l'histoire ou Métier d'historien*. Paris : Armand Colin, 1949.

[6] BROWN, Andrew, ROMAN, Muhammad et DEVEREUX, Barry. « A Systematic Literature Review of Retrieval-Augmented Generation : Techniques, Metrics, and Challenges ». 2025. DOI : 10.48550/arXiv.2508.06401.

[7] BYUN, Courtni, VASICEK, Piper et SEPPI, Kevin. « An Exploration of LLM Citation Accuracy and Relevance ». In : *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Mexico, Mexique : Association for Computational Linguistics, 2024, p. 28-39. DOI : 10.18653/v1/2024.hcinlp-1.3.

[8] CHANDAK, Nikhil, GOEL, Shreya, PRABHU, Aniket, HARDT, Moritz et GEIPING, Jonas. « Answer Matching Outperforms Multiple Choice for Language Model Evaluation ». 2025. DOI : 10.48550/arXiv.2507.02856.

[9] CHEN, Jiawei, LIN, Hongyu, HAN, Xianpei et SUN, Le. « Benchmarking Large Language Models in Retrieval-Augmented Generation ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2024. DOI : 10.1609/aaai.v38i16.29728.

[10] CHERUKURI, Komala Subramanyam, MOSES, Pranav Abishai, SAKATA, Aisa, CHEN, Jiangping et CHEN, Haihua. « Large Language Models for Oral History Understanding with Text Classification and Sentiment Analysis ». 2025. DOI : 10.48550/arXiv.2508.06729.

[11] DA, Nan Z. « The Computational Case against Computational Literary Studies ». In : *Critical Inquiry 45*, no. 3 (2019), p. 601-639. DOI : 10.1086/702594.

[12] DOBSON, James E. « Interpretable Outputs : Criteria for Machine Learning in the Humanities ». In : *Digital Humanities Quarterly 15*, no. 2 (2021). DOI : 10.63744/gqdn7tfwn6r8.

[13] EBERHARD, Katharina. « The Effects of Visualization on Judgment and Decision-Making : A Systematic Literature Review ». In : *Management Review Quarterly 73* (2023), p. 167-214. DOI : 10.1007/s11301-021-00235-8.

[14] JUHASZ, Matyas, DUTIA, Kalyan, FRANKS, Henry, DELAHUNTY, Conor et al. « Responsible Retrieval Augmented Generation for Climate Decision Making from Documents ». 2024. DOI : 10.48550/arXiv.2410.23902.

[15] KARIDI, Danae Pla, CHRYSANTHOPOULOS, Christos et TRIANTAFYLLOU, Ioannis. « Towards a Knowledge-Graph-Driven Retrieval-Augmented Generation for Exploring and Curating Active Archives ». In : *Posters, Demos, Workshops, and Tutorials at SEMANTiCS 2025*. T. 4064. CEUR Workshop Proceedings. Vienne, Autriche, 2025. URL : https://ceur-ws.org/Vol-4064/PD-paper19.pdf.

[16] KELLY, Paul, SCHILD, Jonathan et JAFARI, Amir. « FolkRAG : A Retrieval-Augmented Generation System for Cultural Heritage Materials ». In : *Neural Computing and Applications 37*, no. 24 (2025), p. 20281-20297. DOI : 10.1007/s00521-025-11455-4.

[17] LAMASSÉ Stéphane et Rygiel, Philippe. « Nouvelles frontières de l'historien ». In : *Revue Sciences/Lettres 2* (2014). DOI : 10.4000/rsl.411.

[18] LANGLOIS, Charles-Victor et SEIGNOBOS, Charles. *Introduction aux études historiques*. Paris : Librairie Hachette et Cie, 1898.

[19] LEE, Jeong Ha, ALI, Ghazanfar et HWANG, Jae-In. « A Retrieval-Augmented Generation System for Accurate and Contextual Historical Analysis : AI-Agent for the Annals of the Joseon Dynasty ». In : *Computer Animation & Virtual Worlds 36*, no. 4 (2025). DOI : 10.1002/cav.70048.

[20] LEWIS, Patrick et al. « Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks ». In : *Advances in Neural Information Processing Systems*, sous la dir. d'Hugo LAROCHELLE, Marc' Aurelio RANZATO, Raia HADSELL, Maria-Florina BALCAN et Hsuan-Tien LIN. T. 33. Vancouver, Canada : Curran Associates, Inc., 2020, p. 9459-9474. URL : https://dl.acm.org/doi/abs/10.5555/3495724.3496517.

[21] LIN, Claire, FENG, Bo-Han, CHEN, Xuanjun, YANG, Te-Lun, LEE, Hung-Yi et JANG, Jyh-Shing Roger. « A Preliminary Study of RAG for Taiwanese Historical Archives ». In : *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, sous la dir. de Kai-Wei CHANG, Ke-Han LU, Chih-Kai YANG, Zhi-Rui TAM, Wen-Yu CHANG et Chung-Che WANG. National Taiwan University, Taipei City, Taiwan : Association for Computational Linguistics, 2025, p. 45-62. URL : https://aclanthology.org/2025.rocling-main.6/.

[22] MOUZA, Cédric du, LAMASSÉ, Stéphane et RYGIEL, Philippe. « De l'histoire numérique à l'histoire données ? » In : *Les Cahiers de Framespa 42* (2023). DOI : 10.4000/framespa.14374.

[23] MUDET, Anthony et BAKKALI, Souhail. « Hybrid Retrieval-Augmented Generation for Robust Multilingual Document Question Answering ». 2025. DOI : 10.48550/arXiv.2512.12694.

[24] MULLER, Caroline et CLAVERT, Frédéric. *Écrire l'histoire. Gestes et expériences à l'ère numérique*. Paris : Armand Colin, 2025.

[25] OZAKI, Shintaro, KATO, Yuta, FENG, Siyuan, TOMITA, Masayo, HAYASHI, Kazuki, OBARA, Ryoma, OYAMADA, Masafumi, HAYASHI, Katsuhiko, KAMIGAITO, Hidetaka et WATANABE, Taro. « Understanding the Impact of Confidence in Retrieval Augmented Generation : A Case Study in the Medical Domain ». 2024. DOI : 10.48550/arXiv.2412.20309.

[26] PÄÄKKÖNEN, Juho et YLIKOSKI, Petri. « Humanistic Interpretation and Machine Learning ». In : *Synthese 199* (2021), p. 1461-1497. DOI : 10.1007/s11229-020-02806-w.

[27] PELLET, Aurélien, PEREZ, Julien et PUREN, Marie. « Évaluation automatique du retour à la source dans un contexte historique long et bruité. Application aux débats parlementaires de la Troisième République française ». In : *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, sous la dir. de Vincent CLAVEAU, Nihel KOOLI, Maxime POULAIN et Lorenzo GERARDI. Marseille, France : ATALA & ARIA, 2025, p. 138-150. URL : https://inria.hal.science/hal-05329778.

[28] PELLET, Aurélien, PEREZ, Julien et PUREN, Marie. « Generative Artificial Intelligence and Historical Research : Challenges, Potentials, and Limitations. Application of RAG to French Parliamentary Debates of the Third Republic (1881–1940) ». In : *A Conversation between AI and the Humanities*. Lyon, France, 2024. URL : https://hal.science/hal-04832663.

[29] PELLET, Aurélien, PUREN, Marie et PEREZ, Julien. « HistoriQA-ThirdRepublic : Multi-Hop Question Answering Corpus for Historical Research, Parliamentary Debates from the French Third Republic (1870–1940) ». In : *LREC 2026 : Language Resources and Evaluation Conference*. ELRA Language Resources Association. Palma de Majorque, Espagne, 2026. URL : https://hal.science/hal-05438255.

[30] PIRYANI, Bhawna, MOZAFARI, Jamshid et JATOWT, Adam. « ChroniclingAmericaQA : A Large-Scale Question Answering Dataset Based on Historical American Newspaper Pages ». In : *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '24. Washington, DC, États-Unis : Association for Computing Machinery, 2024, p. 2038-2048. DOI : 10.1145/3626772.3657891.

[31] POUBLANC, Sébastien et MARQUÉ, Nicolas. « Introduction au dossier « Historien·nes et numérique : pratiques et expériences vécues » ». In : *Les Cahiers de Framespa 42* (2023). DOI : 10.4000/framespa.14370.

[32] PROST, Antoine. *Douze Leçons sur l'histoire*. Paris : Éditions du Seuil, 1996.

[33] RAVI, Divya et SINDHGATTA, Renuka. « Exploring Trust and Transparency in Retrieval-Augmented Generation for Domain Experts ». In : *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Yokohama, Japon : Association for Computing Machinery, 2025, 254 :1-254 :7. DOI : 10.1145/3706599.3719985.

[34] REEKEN, Timo von, SALOUS, Mazen et ABDENEBAOUI, Larbi. « Un-trusting the Chat : Designing for Calibrated Trust in Retrieval-Augmented Conversations ». In : *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. CUI '25. New York, États-Unis : Association for Computing Machinery, 2025. DOI : 10.1145/3719160.3737620.

[35] RETTBERG, Jill Walker. « Algorithmic Failure as a Humanities Method ». In : *Big Data & Society 9*, no. 2 (2022), p. 1-6. DOI : 10.1177/20539517221131290.

[36] RIES, Thorsten, DALEN-OSKAM, Karina van et OFFERT, Fabian. « Reproducibility and Explainability in Digital Humanities ». In : *International Journal of Digital Humanities 6* (2024). DOI : 10.1007/s42803-023-00083-w.

[37] SERGEEV, Alexander, GOLOVIZNINA, Valeriya, MELNICHENKO, Mikhail et KOTELNIKOV, Evgeny. « Talking to Data : Designing Smart Assistants for Humanities Databases ». 2025. DOI : 10.48550/arXiv.2506.00986.

[38] TRAN, The Trung, GONZALEZ-GALLARDO, Carlos-Emiliano et DOUCET, Antoine. « Retrieval Augmented Generation for Historical Newspapers ». In : *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries (JCDL '24)*. Hong Kong : Association for Computing Machinery, 2024. DOI : 10.1145/3677389.3702542.

[39] WALLAT, Jonas, HEUSS, Maria, RIJKE, Maarten de et ANAND, Avishek. « Correctness Is Not Faithfulness in Retrieval-Augmented Generation Attributions ». In : *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR '25)*. Padoue, Italie : Association for Computing Machinery, 2025, p. 22-32. DOI : 10.1145/3731120.3744592.

[40] ZHANG, Wan et ZHANG, Jing. « Hallucination Mitigation for Retrieval-Augmented Large Language Models : A Review ». In : *Mathematics 13*, no. 5 (2025), p. 856. DOI : 10.3390/math13050856.

[41] ZHENG, Lianmin et al. « Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena ». In : *Advances in Neural Information Processing Systems*. T. 36. Nouvelle-Orléans, États-Unis : Curran Associates, Inc., 2023. DOI : 10.5555/3666122.3668142.