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

Évaluation des architectures RAG pour la synthèse orientée requête de corpus d'articles d'historiens

Thibault Gautheron[1]
(1) IRIT, 118 Route de Narbonne, 31400 Toulouse, France
thibault.gautheron@irit.fr

RÉSUMÉ

La synthèse multi-documents orientée requête (QFS) constitue un cadre pertinent pour assister les historiens face à l'inflation documentaire. Cependant, l'application des architectures Retrieval-Augmented Generation (RAG) à des corpus d'articles d'historiens pose des défis spécifiques liés à la superposition de temporalités, à l'évolution des concepts et à la pluralité argumentative. À partir d'une baseline RAG appliquée à la revue Le Médiéviste et l'Ordinateur, nous menons une analyse qualitative exploratoire qui met en évidence trois patterns d'erreurs récurrents : biais de généralisation, dérive sémantique et confusion chronologique. En l'absence de résumés de référence, et face à l'inadéquation des métriques classiques pour capturer ces erreurs, nous présentons les premières briques d'un protocole d'évaluation hybride : une grille d'analyse experte structurée en cinq axes, et deux principes de métriques automatiques ciblant spécifiquement la dimension temporelle, une véracité temporelle par alignement source-synthèse et une cohérence chronologique multi-niveaux fondée sur le Tau de Kendall. Ces travaux préliminaires constituent un socle méthodologique pour des recherches ultérieures, notamment vers l'intégration de graphes de connaissances temporels.

ABSTRACT

Evaluation of RAG Architectures for Query-Focused Summarization of Historical Research Article Corpora

Query-Focused Multi-Document Summarization (QFS) provides a relevant framework to assist historians facing documentary inflation. However, applying Retrieval-Augmented Generation (RAG) architectures to corpora of historical research articles raises specific challenges related to overlapping temporalities, conceptual evolution, and argumentative plurality. Based on a baseline RAG system applied to the journal Le Médiéviste et l'Ordinateur, we conduct an exploratory qualitative analysis that highlights three recurring error patterns : generalization bias, semantic drift, and chronological confusion. In the absence of reference summaries, and given the inadequacy of standard metrics for capturing such errors, we present the first building blocks of a hybrid evaluation protocol : an expert-based analytical grid structured along five axes, and two principles for automatic metrics specifically targeting the temporal dimension, a temporal factuality measure based on source–summary alignment and a multi-level chronological coherence measure grounded in Kendall's Tau. These preliminary contributions lay a methodological foundation for further research, in particular toward the integration of temporal knowledge graphs.

MOTS-CLÉS : Humanités numériques, Génération augmentée de récupération, Grands modèles de langage, Synthèse orientée requête.

KEYWORDS: Digital humanities, Retrieval-Augmented Generation, Large Language Models, Query-focused Summarization.

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 14–27.

1 Introduction

La synthèse multi-documents orientée requête (QFS - *Query-Focused Summarization*) est une tâche fondamentale en Traitement Automatique des Langues (TAL) et en Recherche d'Information (RI). Cette tâche consiste à générer un résumé à partir d'un sous-ensemble d'informations pertinent, extrait d'un corpus documentaire sur la base d'une requête (Roy & Kundu, 2023). L'émergence des architectures RAG (*Retrieval-Augmented Generation*) offre un nouveau cadre pour instancier cette tâche. En couplant la puissance générative des grands modèles de langage (LLM) à des systèmes de recherche d'information, le RAG offre une architecture prometteuse permettant de traiter des questions complexes (Gao *et al.*, 2024) tout en limitant les hallucinations (Ji *et al.*, 2023).

Cependant, l'application de ces architectures à des corpus caractérisés par une forte profondeur historique, où les concepts, les sources et les cadres interprétatifs évoluent au fil du temps, constitue un verrou scientifique majeur. Le domaine de l'histoire offre un cas d'étude critique : les textes y sont fortement marqués temporellement et argumentativement. Contrairement à une donnée factuelle stable, un fait historique est souvent une position située dans un débat (Poublanc & Charles, 2025).

Les modèles génériques actuels peinent à traiter cette complexité historiographique pour plusieurs raisons.

Tout d'abord, ils souffrent d'une faible prise en compte des dérives sémantiques. Dans la majorité des pipelines RAG standards basés sur des embeddings génériques, la recherche d'information repose sur une similarité sémantique statique, qui peine à distinguer les concepts dont le sens évolue au fil du temps (Tahmasebi *et al.*, 2019) (Hegde *et al.*, 2025). Cela risque d'entraîner des anachronismes conceptuels majeurs (ex : projeter un sens moderne sur un terme ancien).

De plus, les LLM utilisés pour la génération de la synthèse tendent intrinsèquement au lissage du consensus. Entraînés pour maximiser la vraisemblance statistique, ils tendent à gommer les controverses et les positions minoritaires pour produire une narration unifiée et fluide (Bender *et al.*, 2021; Tamkin *et al.*, 2021). Or, en histoire, la valeur d'un document réside souvent dans sa singularité argumentative. Cette tendance au lissage entraîne une perte d'information, transformant des hypothèses débattues en vérités générales.

Enfin, les modèles de langage souffrent de difficultés par rapport à la représentation du temps, et au raisonnement temporel (Fatemi *et al.*, 2024) (Wallat *et al.*, 2025). Or, les textes d'historiens sont fondamentalement structurés par une chronologie complexe. Ils superposent le temps de l'objet d'étude (événements historiques), le temps de la production des sources, et le temps de l'écriture par l'historien. Cette difficulté des modèles à modéliser le temps peut mener à des synthèses anachroniques, mélangeant les faits relevant de différentes temporalités. Si ces trois limites, dérive sémantique, lissage du consensus, et faiblesse temporelle concourent toutes à dégrader la fidélité historique des synthèses, c'est sur la dimension temporelle que nous concentrons les premières propositions méthodologiques de cet article, à la fois en raison de sa centralité pour les corpus d'historiens et de l'absence de métriques automatiques dédiées dans la littérature.

Les travaux présentés dans cet article s'intéressent à la modélisation de synthèses documentaires adaptées aux besoins experts des historiens. Les historiens sont actuellement engagés dans une réflexion épistémologique sur les fondements méthodologiques de la discipline. Il s'agit d'un changement de paradigme : auparavant, les articles scientifiques d'historiens ne mettaient l'accent que sur les seuls résultats de recherche ; aujourd'hui, il est nécessaire de révéler les approches instrumentées et méthodologiques utilisées pour produire ces connaissances afin de prendre conscience que l'histoire n'est

Requêtes	Dimensions requête						Dimensions synthèse					
	Objet	Terrains	Epoques	Données	Outils	Méthodes	Objet	Terrains	Epoques	Données	Outils	Méthodes
Quels travaux ont étudié les transformations du territoire correspondant à l'actuelle Guinée entre 1890 et 1960 ?	×	×	×					×	×			
Quels sont les objets d'étude concernant l'actuelle Guinée entre 1890 et 1960 ?	×	×	×				×	×	×			
Quels sont les jeux de données utilisés pour les recherches concernant l'actuelle Guinée entre 1890 et 1960 ?		×	×	×				×	×	×		
Quelles transformations méthodologiques sont identifiées entre les années 1980 et 2000 dans les études médiévales assistées par ordinateur ?			×			×			×		×	×
Quels sont les premiers usages de l'informatique documentaire dans les laboratoires d'histoire médiévale ?					×	×			×		×	×

TABLE 1 – Typologie des requêtes et dimensions associées à la requête et à la synthèse attendue

pas qu'une science humaine : c'est aussi une science instrumentée. Face à l'inflation documentaire, l'enjeu n'est plus seulement d'accéder à l'information, mais de réussir à identifier, extraire et mettre en perspective ces méthodes de construction du savoir. Contrairement à une recherche d'information classique, une requête d'historien ne cherche donc pas une réponse unique, mais une mise en perspective croisant plusieurs dimensions d'analyse : l'objet d'étude (la thématique historique abordée), les terrains (l'aspect spatial ou géographique), les époques (la temporalité des évenements), les jeux de données (sources documentaires exploitées), les outils (l'instrumentation technique) et les méthodes d'analyse employées. Ces dimensions structurent l'argumentation attendue par l'historien. Une même thématique peut ainsi être abordée selon des terrains différents, à partir de jeux de données distincts ou via des méthodologies contrastées (quantitatives, qualitatives, computationnelles), produisant des interprétations hétérogènes.

Comme l'illustre le **Tableau 1**, composés de requêtes synthétiques inspirées de discussions avec des historiens, une particularité supplémentaire est qu'il peut exister un écart entre les dimensions explicitement mentionnées dans la requête de l'utilisateur et les dimensions attendues dans la synthèse.

L'objectif de cette recherche est de dépasser les limitations des modèles génériques pour la génération de synthèse multi-documents orientée-requête sur des corpus de travaux d'historiens. Notre question de recherche est la suivante : *Comment orienter et évaluer un système RAG pour générer une synthèse qui respecte les dimensions attendues par l'historien et qui réponde précisément, de manière fidèle aux sources, à sa requête ?*

Dans cet article, nous présentons les résultats de nos premiers travaux exploratoires, dont le focus se resserre progressivement sur la dimension temporelle, jugée critique pour ce type de corpus :

— **Analyse qualitative des synthèses :** Nous établissons une *baseline* RAG sur un corpus de référence que nous avons choisi et proposons une analyse qualitative des erreurs rencontrées.
— **Protocole d'évaluation hybride :** Face à l'inadéquation des métriques classiques pour mesurer la véracité historique, nous proposons une base de protocole d'évaluation couplant

une grille d'analyse qualitative et les principes théoriques de deux métriques automatiques ciblant spécifiquement la dimension temporelle.

2 Analyse critique de l'état de l'art

Dans cette section, nous passons en revue l'état de l'art sur la synthèse multi-documents orientée requête (QFS), avec un focus sur les architectures RAG et les approches de synthèse temporelles. Nous discutons des limites de ces méthodes pour traiter des corpus d'articles d'historiens, en mettant en évidence les défis liés à la temporalité et à l'évolution des concepts.

2.1 QFS et RAG

La synthèse multi-documents orientée requête (*Query-Focused Summarization*, QFS) consiste à produire un résumé concis répondant à un besoin informationnel explicite, formulé sous la forme d'une requête (Dang, 2005; Nenkova & McKeown, 2012). Contrairement à la synthèse de documents générique, la QFS impose de sélectionner et d'organiser uniquement les informations pertinentes pour la requête, tout en maintenant la cohérence globale du résumé (Maybury & Mani, 2001). Cette tâche est rendue plus difficile si effectuée sur des multiples documents, le résumé doit prendre en compte la redondance de l'information et les possibles divergences entre les sources (Nenkova & McKeown, 2012).

Historiquement, la QFS a été structurée par les campagnes DUC, qui ont favorisé le développement de méthodes extractives ou hybrides fondées sur des modèles de pertinence et de couverture (Dang, 2005). L'essor des approches neuronales a ensuite permis de développer des modèles plus abstractifs, capables de paraphraser, mais sensibles aux biais de génération (See *et al.*, 2017).

Dans ce contexte, le paradigme *Retrieval-Augmented Generation* (RAG) (Lewis *et al.*, 2021) s'est imposé comme une architecture de référence pour ancrer la génération dans des documents externes. Le RAG couple un module de recherche d'information et un modèle génératif, afin de produire une réponse ou une synthèse à partir de passages récupérés (Gao *et al.*, 2024). Dans le cadre de la QFS, le module de recherche (*retriever*) récupère un contexte supposé pertinent par rapport à la requête, tandis que le module génératif réalise la structuration et la compression des informations pour générer la synthèse (Zhong *et al.*, 2021).

De plus, dans des corpus constitués d'articles d'historiens, la pertinence du module de recherche ne doit pas se réduire à une similarité sémantique locale statique (Tahmasebi *et al.*, 2019). Les textes d'historiens mobilisent des terminologies évolutives, des cadres interprétatifs concurrents et des positions argumentatives situées (Genet, 1986). Cela suggère que l'adaptation du RAG à la QFS sur des travaux d'historiens requiert des mécanismes sensibles au contexte temporel et à la diversité des points de vue.

Enfin, plusieurs travaux soulignent que la qualité du *retrieval* ne garantit pas une génération fidèle : même en présence de portions de textes correctes, le modèle peut omettre des informations critiques (Liu *et al.*, 2023a) ou produire des attributions erronées (Gao *et al.*, 2023)(Ji *et al.*, 2023). Ces limites sont particulièrement importantes pour la synthèse multi-documents, où l'objectif est de restituer un ensemble de contributions potentiellement hétérogènes (Kryściński *et al.*, 2019).

2.2 Synthèse temporelle

La prise en compte de la dimension temporelle en synthèse automatique a donné lieu à des travaux regroupés sous les notions de *Temporal Summarization* et *Time-Aware Summarization*, dont l'objectif est de produire des synthèses structurées par une chronologie (Aslam *et al.*, 2015). Dans les contextes les plus étudiés, en particulier les flux d'actualités, la tâche vise souvent à générer un résumé évolutif ou une *timeline* décrivant l'évolution d'un événement (Yan *et al.*, 2011).

Les approches classiques reposent généralement sur l'extraction d'expressions temporelles et d'événements, leur normalisation, puis leur ordonnancement pour guider la sélection et l'organisation du résumé (Lloret & Palomar, 2012). Plusieurs travaux introduisent ainsi des structures intermédiaires comme des chaînes temporelles, des graphes d'événements, ou des *timelines* afin de contraindre la cohérence chronologique (Do *et al.*, 2012).

Cependant, ces travaux reposent le plus souvent sur une hypothèse : le temps est modélisé comme une séquence linéaire d'événements factuels, associés à des dates explicites (Gholipour Ghalandari *et al.*, 2020). Cette hypothèse est adaptée aux actualités, mais devient insuffisante dans des corpus d'articles d'historiens, où les documents superposent le temps de l'objet d'étude (ex : évènement au XIIIème siècle), le temps des sources (ex : charte du XVème siècle), et le temps de l'écriture scientifique (ex : article de 1985) (Genet, 1986). Dans ce contexte, une synthèse correcte doit non seulement ordonner des faits, mais aussi maintenir leur attribution à la bonne temporalité.

En parallèle, la littérature récente montre que les modèles de langue peinent à manipuler des relations temporelles explicites (Xiong *et al.*, 2024), à maintenir des contraintes chronologiques sur des contextes longs (Zhang *et al.*, 2024), et à éviter les inversions d'ordre entre événements (Fatemi *et al.*, 2024). Ces limites sont amplifiées en synthèse multi-documents, où la fusion d'informations issues de périodes différentes augmente le risque d'anachronismes (Wallat *et al.*, 2025).

De plus, la temporalité en histoire ne se réduit pas à l'ordre chronologique : elle est liée à l'évolution des concepts au fil du temps. Ce phénomène, étudié sous le nom de changement sémantique ou dérive lexicale, a donné lieu à une littérature importante en linguistique computationnelle, notamment via les *embeddings diachroniques* (Tahmasebi *et al.*, 2019; Kutuzov & Giulianelli, 2020). Or, cette dimension reste rarement intégrée dans les travaux de synthèse temporelle, ce qui peut conduire à des synthèses chronologiquement cohérentes mais conceptuellement anachroniques.

Ces constats suggèrent que les approches de synthèse temporelle existantes, majoritairement conçues pour des flux d'actualités, doivent être adaptées pour traiter des corpus constitués d'articles historiques, en intégrant à la fois les multiples temporalités et l'évolution des concepts.

2.3 Evaluation des synthèses

L'évaluation de la qualité des synthèses repose traditionnellement sur des comparaisons avec des résumés de référence. Historiquement conçues pour l'évaluation de la traduction automatique ou du résumé de texte, les métriques de surface comme ROUGE (Lin, 2004) et BLEU (Papineni *et al.*, 2001) ont été largement adoptées pour évaluer les tâches de QFS dans les campagnes d'évaluation DUC (Dang, 2005). Cependant, cette approche par chevauchement de n-grammes comporte un biais majeur pour l'analyse historique : elle favorise la correspondance lexicale au détriment de la structure causale et chronologique. Par exemple, une métrique basée sur les unigrammes attribuera un score de similarité parfait entre la phrase "L'outil a précédé la méthode" et "La méthode a précédé l'outil",

alors que le sens historique s'en trouve inversé.

Des métriques sémantiques ont été proposées pour dépasser ces limites, notamment BERTScore (Zhang *et al.*, 2020) ou MoverScore (Zhao *et al.*, 2019), qui comparent les textes via des représentations contextuelles. Si ces mesures capturent mieux la paraphrase et la synonymie, elles restent insuffisantes pour juger de la véracité factuelle et de la cohérence temporelle (Pagnoni *et al.*, 2021). Dans un corpus d'historiens, une synthèse générée peut être sémantiquement très proche de la référence tout en contenant des anachronismes ou des erreurs d'attribution d'une méthode à une mauvaise époque.

Par ailleurs, la production de résumés de référence est particulièrement coûteuse, car elle requiert une expertise fine du corpus documentaire. Cette difficulté a favorisé l'émergence de métriques sans référence (*reference-free*) orientées vers la vérification factuelle. Les approches basées sur l'inférence en langage naturel (NLI) ou la génération de questions (*QA-based*), telles que QuestEval (Scialom *et al.*, 2021), QAGS (Wang *et al.*, 2020) ou SummaC (Laban *et al.*, 2021), ainsi que les récents cadres d'évaluation utilisant les LLM comme juges (*LLM-as-a-judge*) — à l'instar de G-Eval (Liu *et al.*, 2023b) et ARES (Saad-Falcon *et al.*, 2024) — en sont des exemples représentatifs.

Ces approches évaluent la cohérence factuelle en vérifiant si les affirmations de la synthèse peuvent être déduites des documents sources. Si elles détectent efficacement certaines hallucinations comme les entités inventées ou les faits inexistants, elles restent généralement insensibles aux contraintes diachroniques (Bajpai *et al.*, 2024). En effet, elles évaluent la présence d'un fait indépendamment de sa validité temporelle croisée (Lee *et al.*, 2025). Or, une affirmation peut être vraie dans une période mais fausse dans une autre (ex : usage d'un outil, terminologie, disponibilité d'un jeu de données).

Face à ces limites, il n'existe pas à notre connaissance de protocole automatisé permettant d'évaluer l'adéquation d'une synthèse avec la mise en perspective attendue par les historiens. Ce constat justifie notre proposition d'un protocole d'évaluation hybride dans la suite de cet article, associant une grille d'analyse humaine ciblée et des métriques automatiques.

3 Analyse de premiers résultats

Afin de tester la capacité des modèles de langage à gérer les spécificités et les difficultés des documents historiques, nous avons établi une baseline RAG sur notre corpus, constitué des articles de la revue Le Médiéviste et l'Ordinateur, revue française consacrée aux méthodes numériques en histoire.

3.1 Protocole expérimental

Corpus. Les expérimentations ont été menées sur l'intégralité de la revue *"Le Médiéviste et l'Ordinateur"*. Ce corpus, publié entre 1979 et 2003, est constitué de 45 numéros et 733 articles. Il présente une forte hétérogénéité structurelle, les articles étant répartis entre éditoriaux, articles de fond, brèves techniques et comptes-rendus de séminaires. Il y a par exemple 39 éditoriaux, 394 articles de fond ou 123 comptes-rendus dans le corpus.

Architecture. Notre architecture de référence se compose du modèle de recherche `BGEm3` pour la partie *retrieval* et du modèle de langage `Qwen-3 7b` pour la génération. Nous segmentons les

articles en passages de 1200 tokens avec un recouvrement de 100 tokens. Pour chaque requête, nous récupérons les k = 10 passages les plus similaires selon BGEm3. Ces passages sont concaténés pour former le contexte d'entrée du modèle génératif. Dans nos expériences, nous n'appliquons pas de contrainte supplémentaire de compression du contexte, afin d'observer les erreurs d'une baseline RAG standard. Le prompt utilisé pour la génération est le suivant :

> **Rôle assigné :**
> Tu es un historien spécialisé dans les humanités numériques. Tu analyses un corpus d'articles d'historiens issu de la revue *Le Médiéviste et l'Ordinateur* (1979-2003). Tu dois répondre à la requête suivante le plus précisement possible en te basant sur le contexte fourni.
>
> **Requête :** `{requete}`
>
> **Contraintes strictes de génération :**
> — **Absence d'anachronismes :** N'utilise pas de vocabulaire moderne (ex. : "IA", "Big Data") pour décrire les outils de l'époque, sauf si le terme figure explicitement dans le texte source. Emploie la terminologie historique des auteurs (ex. : "sérielle", "lexicométrie", "automatique").
> — **Posture épistémologique :** Distingue rigoureusement le discours des auteurs (leurs espoirs, croyances ou militantisme) de la réalité technique avérée. Si un auteur de 1985 espère qu'une machine va tout changer, rapporte cette affirmation comme une croyance située dans son époque, et non comme un fait.
> — **Attribution des sources :** Cite systématiquement les auteurs (ex. : "Comme le souligne Zysberg...") et identifie la nature du document (ex. : "Dans son éditorial de 1990...").
> **Contexte fourni :** `{input_information}`

Jeu de requêtes. Pour évaluer les synthèses, nous avons commencé par générer un ensemble de 50 requêtes représentatives des interrogations historiennes. Ces requêtes ont été affinées et validées avec un historien pour garantir leur pertinence. À l'issue de cette étape, 5 requêtes ont été écartées car elles portaient sur des thématiques absentes du corpus ou étaient redondantes. Le jeu final compte ainsi 45 requêtes validées.

Afin de structurer ce jeu de requêtes, nous l'avons classifié selon trois axes d'interrogation. Notre catégorisation assume un recouvrement des classes, une requête complexe peut croiser plusieurs axes simultanément.

- **L'axe historiographique et épistémologique** (28 occurrences) : discours, réseaux d'acteurs et débats.
- **L'axe méthodologique et technique** (30 occurrences) : instruments, logiciels et méthodes.
- **L'axe spatial, temporel et thématique** (24 occurrences) : modélisation des objets, terrains et époques.

3.2 Analyse qualitative

Pour chaque requête de notre jeu d'évaluation, nous avons généré une synthèse via notre architecture RAG. Pour cette étude exploratoire, nous avons choisi d'analyser en détail une synthèse dont la requête mobilise les trois axes de notre classification. L'analyse s'est portée sur la requête suivante : *« Comment les historiens décrivent-ils l'impact des outils informatiques sur leur pratique de la recherche historique ? »*. Cette interrogation est particulièrement riche car elle croise trois dimensions fondamentales de notre typologie (cf. Tableau 1), la dimension **Outils** (l'informatique, ses instruments matériels et logiciels), la dimension **Méthodes** (la pratique de la recherche et les cadres théoriques) et la dimension **Époques**, une dimension attendue puisqu'on demande de restituer l'évolution de cet

impact sur le temps long de la revue (1979-2003).

Cette analyse approfondie sur une seule requête permet de mettre en évidence des patterns d'erreurs qualitativement riches, mais ne permet pas à elle seule d'établir leur fréquence ni leur généralisation à l'ensemble du jeu d'évaluation. L'application systématique de la grille d'analyse aux 45 requêtes constitue l'une des prochaines étapes de notre travail.

Extrait de la synthèse RAG :
"Les médiévistes, dans les textes du Médiéviste et l'Ordinateur (1979–2000), décrivent l'impact des outils informatiques sur leur pratique de la recherche historique comme un processus d'ouverture, d'optimisation et de transformation méthodologique, mais toujours marqué par des espoirs ambigus, des craintes structurelles et une conscience du fossé entre la promesse et la réalité technique. **Le discours des auteurs est militant et optimiste dès les premières années**. **L'éditorial du n°1 (1979)** s'inscrit dans une volonté d'« évangélisation » : il cherche à convaincre les réticents que l'informatique peut porter ses fruits. Ce n'est pas seulement un outil, mais un intermédiaire méthodologique qui permet de réorganiser les sources, de les **gérer de manière automatisée** et de faciliter les recherches. L'accent est mis sur des usages concrets : la **gestion automatisée** de fichiers documentaires, le découpage et le codage des données, voire la mise en œuvre d'outils de lexicologie pour l'analyse des textes. Les auteurs voient dans l'informatique un **outil de libération** du chercheur des tâches répétitives : **« débarassés de la plus grande partie des travaux répétitifs » (Bourlet, 1993)**. Leur espoir est d'accélérer les processus de recherche, notamment dans les domaines de l'archéologie ou de l'analyse des manuscrits, où les données sont volumineuses. L'image du réseau est particulièrement valorisée : il permet de reconstituer des groupes de recherche dispersés, d'organiser des échanges entre chercheurs et de faciliter la diffusion de l'information. **À partir des années 1980**, [...]"

L'analyse critique de cet extrait, effectuée avec un historien spécialiste de la revue, nous a révèlé les distorsions suivantes :

Biais de généralisation : Le modèle affirme que **"le discours des auteurs est militant et optimiste dès les premières années"**, en s'appuyant sur l'éditorial du premier numéro. Le qualificatif 'militant' est ici fidèle à la source : l'éditorial de 1979 adopte effectivement un registre d'engagement explicite. Un éditorial est un texte d'intention, rédigé par les fondateurs de la revue pour promouvoir une vision (ici, encourager les historiens à utiliser l'informatique). Or, le modèle généralise cette posture militante à l'ensemble des auteurs de l'époque. Pourtant, la lecture des articles de recherche publiés en 1979 montre une réalité opposée : les historiens de terrain y décrivent des difficultés techniques lourdes (pannes, complexité des cartes perforées) et de profonds doutes méthodologiques. En accordant un poids important à l'éditorial, le modèle lisse les controverses et fausse la réalité historique en généralisant l'optimisme des fondateurs de la revue.

Dérive sémantique : L'expression **"gérer de manière automatisée"**, appliquée au contexte de 1979, relève d'une projection moderne. À cette époque, la "gestion" impliquait un codage manuel rigide sur bordereaux et des traitements par lots différés. En utilisant un vocabulaire évoquant la fluidité des systèmes actuels, le modèle fausse la compréhension de la méthodologie historienne de l'époque.

Dérive temporelle : La confusion chronologique est ici flagrante. Le paragraphe débute en analysant **les premières années de la revue (1979)**, puis justifie le concept **"d'outil de libération"** par une citation de Bourlet datant de **1993**. Attribuer le sentiment de "libération" de 1993 aux pionniers de 1979 constitue un contresens historique majeur. Le modèle continue après cela sa chronologie en

cherchant le point de vue des auteurs des **années 80**. En cherchant à valider son récit pour garantir la fluidité de sa réponse, le modèle a extrait une citation sémantiquement pertinente, mais n'a pas vérifié sa validité temporelle, qui n'est pas applicable sur la période de 1979.

Cet exemple montre que le modèle privilégie la cohérence du texte au détriment de la chronologie. En voulant créer une réponse fluide, il mélange des périodes distinctes et efface les évolutions du discours historien. Ces biais montrent que, sans ajustements spécifiques, le RAG a des difficultés à restituer fidèlement les temporalités et la diversité des discours historiques.

4 Propositions

Pour évaluer ces synthèses, nous ne pouvons pas utiliser de métriques nécessitant des résumés de références. Plus généralement, les corpus de travaux d'historiens ne disposent pas de résumés *Gold Standard*. L'application des métriques standards en synthèse d'information (ROUGE, BLEU) est donc impossible. Créer ces références à grande échelle s'avère également irréalisable en pratique. D'une part, la rédaction d'une synthèse exige la lecture croisée d'articles denses, ce qui impose un coût cognitif et temporel prohibitif tout en requérant une expertise disciplinaire pointue. D'autre part, l'interprétation des sources historiques admet souvent plusieurs synthèses valides selon l'angle d'analyse choisi, démultipliant encore plus l'effort requis. Si les métriques *référence-free* existantes, telles que BERTScore ou QuestEval, offrent des pistes pour la cohérence sémantique et factuelle, elles ne couvrent pas la dimension temporelle, qui est critique dans le domaine historien.

Nous proposons donc une approche hybride, commençant par une grille manuelle pour obtenir des scores humains de référence, puis développant des métriques automatiques que nous viendrons corréler avec notre grille.

4.1 Grille d'évaluation

Pour faire nos premières évaluations, nous avons créé une grille d'évaluation manuelle structurée en 5 axes fondamentaux :

1. **Temporalisation (Dimension *Époques*) :** La synthèse est-elle ancrée dans un cadre chronologique explicite et exact ?
2. **Couverture dimensionnelle (Dimensions *Outils, Méthodes, Objets, Terrains, Données*) :** Le résumé traite-t-il tous les axes analytiques exigés par la requête ?
3. **Discours :** Le modèle respecte-t-il le statut des textes (usage de la nuance et du conditionnel, capacité à distinguer une croyance d'un fait technique) ?
4. **Cohérence :** Les paragraphes s'enchaînent-ils logiquement, sans contradictions internes et sans créer de fausses associations méthodologiques ?
5. **Ancrage :** Les informations, citations et entités nommées proviennent-elles réellement du corpus et sont-elles correctement sourcées ou attribuées à leurs auteurs légitimes ?

Chaque axe est sous-divisé en plusieurs questions. Chaque question est évaluée sur une échelle de Likert à 4 points (de 0 : "résultat très insatisfaisant" à 3 : "résultat très satisfaisant"). Cette grille appliquée aux synthèses nous permettra ultérieurement de procéder à des calculs de corrélation entre ces scores qualitatifs et les métriques automatiques que nous développons.

4.2 Métriques temporelles

Nos travaux d'automatisation se focalisent prioritairement sur la dimension temporelle. Si la littérature en *Temporal Summarization* propose des structures intermédiaires (chaînes temporelles, timelines) pour contraindre la cohérence chronologique (Aslam *et al.*, 2015; Do *et al.*, 2012), ces approches supposent une chronologie unique et linéaire et ne permettent pas d'évaluer simultanément la véracité d'une datation et la cohérence d'une narration qui superpose plusieurs strates temporelles. Les métriques de cohérence factuelle existantes (Scialom *et al.*, 2021; Laban *et al.*, 2021), de leur côté, ne contraignent pas l'attribution temporelle des faits. Les deux principes de métriques ci-dessous visent à combler ces lacunes ; leur implémentation est en cours.

Véracité temporelle. Pour évaluer la précision des dates liées aux entités nommées, nous voulons adopter une approche de vérification par alignement source-synthèse. Le principe consiste à apparier sémantiquement chaque phrase générée avec son passage de référence dans le corpus. Cela permet de contrôler que chaque mention temporelle est correctement attribuée à son événement ou son entité, garantissant ainsi la véracité historique des faits rapportés.

Cependant, une comparaison littérale montre rapidement ses limites face à la richesse des expressions temporelles historiques. Il est fréquent qu'une synthèse généralise une date précise (e.g., source : "1250") par une période (résumé : "XIII$^{\text{e}}$ siècle" ou "Moyen Âge central"). Pour qu'une métrique automatique ne pénalise pas cette généralisation légitime, il est nécessaire de construire un référentiel temporel.

Nous prévoyons donc la construction d'un graphe de connaissances temporel qui projette chaque mention vers un intervalle standardisé $[t_{\text{début}}, t_{\text{fin}}]$. La vérification se réduit alors à une inclusion d'intervalles ($1250 \in [1201, 1300]$). Le score global agrège ces vérifications à l'échelle de la synthèse.

Chronologie multi-niveaux. Pour mesurer la chronologie et vérifier une évolution cohérente des temporalités, nous allons créer une métrique basée sur le principe du Tau de Kendall, qui compare deux ordonnancements d'un même ensemble d'éléments en comptant les paires concordantes et discordantes (valeurs dans $[-1, +1]$). Nous comparons l'ordre des dates dans la synthèse à leur ordre dans les passages sources alignés.

L'originalité tient à la prise en compte des trois strates temporelles propres aux textes d'historiens, le temps de l'objet d'étude, le temps des sources et le temps de l'écriture scientifique, qui peuvent être entremêlées dans le texte sans constituer une erreur.

Nous proposons donc de séparer les dates par strate et de calculer un Tau de Kendall sur chacune, avant d'agréger les résultats à l'échelle de la synthèse. La séparation des strates est le point le plus délicat : nous explorons diverses méthodes, comme des règles heuristiques, une combinaison de reconnaissance d'entités temporelles et de classification supervisée, ou du LLM-as-a-Judge. Ces diverses méthodes devront elles-mêmes faire l'objet d'une évaluation pour évaluer leur fiabilité.

Pour une synthèse contenant la séquence $[1250, 1979, 1420, 1300, 1993, 1450]$, la séparation produit trois listes : objet ($[1250, 1300]$), sources ($[1420, 1450]$), écriture ($[1979, 1993]$). Si l'ordre d'apparition de ces dates dans les sources alignées est identique, les trois Tau valent $+1$. Une synthèse produisant $[1300, 1250, 1420, 1450, 1979, 1993]$ obtiendrait en revanche un Tau de -1 sur la strate « objet » tout en maintenant $+1$ sur les deux autres, signalant une inversion locale au sein d'un texte globalement cohérent.

Ces deux métriques ne couvrent pas l'évolution sémantique des concepts. L'intégration d'embeddings diachroniques (Tahmasebi *et al.*, 2019) pour détecter les anachronismes conceptuels constitue une perspective intéressante, laissée à des travaux ultérieurs.

5 Conclusion

L'application de la tâche de QFS aux corpus d'articles d'historiens ouvre des perspectives stimulantes, mais révèle également les limites des architectures RAG standard sur ce type de documents. Nos premières expérimentations sur la revue *Le Médiéviste et l'Ordinateur* montrent que les modèles RAG peinent à maintenir la rigueur chronologique et tendent à lisser les évolutions conceptuelles, tout en générant des anachronismes et des interprétations idéalisées des pratiques historiques.

Pour répondre à ces défis, nous avons posé les bases d'une méthodologie d'évaluation, croisant l'expertise humaine avec le développement de métriques automatiques sensibles à la temporalité et aux changements conceptuels.

Deux limites invitent à prolonger ce travail : d'une part, étendre l'analyse qualitative à l'ensemble des 45 requêtes pour quantifier les patterns d'erreurs dégagés dans cette étude et d'autre part, diversifier les couples retrieval/génération au-delà de BGE-m3 et Qwen-3 7b, afin de mieux isoler les erreurs imputables au corpus de celles qui relèvent des modèles eux-mêmes.

Dans la continuité de cette étude, nos travaux futurs viseront également à expérimenter l'intégration de graphes de connaissances temporels afin de contraindre le modèle sur les chronologies et les évolutions conceptuelles. Cette approche hybride constitue une piste prometteuse pour fournir au modèle un guidage et produire des synthèses robustes, temporellement fidèles et respectueuses des requêtes de l'historien.

Références

ASLAM J., DIAZ F., EKSTRAND-ABUEG M., MCCREADIE R., PAVLU V. & SAKAI T. (2015). Trec 2014 temporal summarization track overview.

BAJPAI A., GOYAL A., ANWER A. & CHAKRABORTY T. (2024). Temporally consistent factuality probing for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 15864–15881 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.887.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

DANG H. T. (2005). Overview of duc 2005.

DO Q., LU W. & ROTH D. (2012). Joint inference for event timeline construction. In J. TSUJII, J. HENDERSON & M. PAŞCA, Éds., *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 677–687, Jeju Island, Korea : Association for Computational Linguistics.

FATEMI B., KAZEMI M., TSITSULIN A., MALKAN K., YIM J., PALOWITCH J., SEO S., HALCROW J. & PEROZZI B. (2024). Test of time : A benchmark for evaluating llms on temporal reasoning.

GAO T., YEN H., YU J. & CHEN D. (2023). Enabling large language models to generate text with citations. In *Conference on Empirical Methods in Natural Language Processing*.

GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., WANG M. & WANG H. (2024). Retrieval-augmented generation for large language models : A survey.

GENET J.-P. (1986). Histoire, informatique, mesure. *Histoire Mesure*, **1**(1), 7–18. DOI : 10.3406/hism.1986.904.

GHOLIPOUR GHALANDARI D., HOKAMP C., PHAM N. T., GLOVER J. & IFRIM G. (2020). A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1302–1308, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.120.

HEGDE N., PAUL S., JOEL-FREY L., BRACK M., KERSTING K., MUNDT M. & SCHRAMOWSKI P. (2025). Chronoberg : Capturing language evolution and temporal awareness in foundation models.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y., CHEN D., DAI W., CHAN H. S., MADOTTO A. & FUNG P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, **55**(12), 1–38. arXiv :2202.03629 [cs], DOI : 10.1145/3571730.

KRYŚCIŃSKI W., MCCANN B., XIONG C. & SOCHER R. (2019). Evaluating the factual consistency of abstractive text summarization.

KUTUZOV A. & GIULIANELLI M. (2020). UiO-UvA at SemEval-2020 task 1 : Contextualised embeddings for lexical semantic change detection. In A. HERBELOT, X. ZHU, A. PALMER, N. SCHNEIDER, J. MAY & E. SHUTOVA, Éds., *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, p. 126–134, Barcelona (online) : International Committee for Computational Linguistics. DOI : 10.18653/v1/2020.semeval-1.14.

LABAN P., SCHNABEL T., BENNETT P. N. & HEARST M. A. (2021). Summac : Re-visiting nli-based models for inconsistency detection in summarization.

LEE D., JUNG H. & CHOI Y. S. (2025). Mind the link : Discourse link-aware hallucination detection in summarization. *Applied Sciences*, **15**(19). DOI : 10.3390/app151910506.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv :2005.11401 [cs], DOI : 10.48550/arXiv.2005.11401.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LIU N. F., LIN K., HEWITT J., PARANJAPE A., BEVILACQUA M., PETRONI F. & LIANG P. (2023a). Lost in the middle : How language models use long contexts.

LIU Y., ITER D., XU Y., WANG S., XU R. & ZHU C. (2023b). G-eval : NLG evaluation using gpt-4 with better human alignment. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 2511–2522, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.153.

LLORET E. & PALOMAR M. (2012). Text summarisation in progress : A literature review. *Artif. Intell. Rev.*, **37**, 1–41. DOI : 10.1007/s10462-011-9216-z.

MAYBURY M. T. & MANI I. (2001). Automatic summarization. In *ACL/EACL*, volume 1.

NENKOVA A. & MCKEOWN K. (2012). A survey of text summarization techniques. In *Mining Text Data*.

PAGNONI A., BALACHANDRAN V. & TSVETKOV Y. (2021). Understanding factuality in abstractive summarization with frank : A benchmark for factuality metrics.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2001). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, p. 311, Philadelphia, Pennsylvania : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

POUBLANC S. & CHARLES W. (2025). Parler d'informations, de données, de data... en histoire et en informatique : réflexions interdisciplinaires à partir du projet ANR ObARDI. In *Workshop Qualité et incertitude des données historique*, Toulouse, France : FRAMESPA and Projets Time Machine. HAL : hal-05485236.

ROY P. & KUNDU S. (2023). Review on query-focused multi-document summarization (qmds) with comparative analysis. *ACM Comput. Surv.*, **56**(1). DOI : 10.1145/3597299.

SAAD-FALCON J., KHATTAB O., POTTS C. & ZAHARIA M. (2024). Ares : An automated evaluation framework for retrieval-augmented generation systems.

SCIALOM T., DRAY P.-A., GALLINARI P., LAMPRIER S., PIWOWARSKI B., STAIANO J. & WANG A. (2021). QuestEval : Summarization Asks for Fact-based Evaluation. arXiv :2103.12693 [cs], DOI : 10.48550/arXiv.2103.12693.

SEE A., LIU P. J. & MANNING C. D. (2017). Get to the point : Summarization with pointer-generator networks.

TAHMASEBI N., BORIN L. & JATOWT A. (2019). Survey of computational approaches to lexical semantic change.

TAMKIN A., BRUNDAGE M., CLARK J. & GANGULI D. (2021). Understanding the capabilities, limitations, and societal impact of large language models.

WALLAT J., ABDALLAH A., JATOWT A. & ANAND A. (2025). A study into investigating temporal robustness of LLMs. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2025*, p. 15685–15705, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-acl.810.

WANG A., CHO K. & LEWIS M. (2020). Asking and answering questions to evaluate the factual consistency of summaries.

XIONG S., PAYANI A., KOMPELLA R. & FEKRI F. (2024). Large language models can learn temporal reasoning.

YAN R., KONG L., HUANG C., WAN X., LI X. & ZHANG Y. (2011). Timeline generation through evolutionary trans-temporal summarization. In R. BARZILAY & M. JOHNSON, Éds., *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 433–443, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating Text Generation with BERT. arXiv :1904.09675 [cs], DOI : 10.48550/arXiv.1904.09675.

ZHANG Z., CAO Y., YE C., MA Y., LIAO L. & CHUA T.-S. (2024). Analyzing temporal complex events with large language models ? a benchmark towards temporal, long context understanding.

ZHAO W., PEYRARD M., LIU F., GAO Y., MEYER C. M. & EGER S. (2019). MoverScore : Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 563–578, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1053.

ZHONG M., YIN D., YU T., ZAIDI A., MUTUMA M., JHA R., AWADALLAH A. H., CELIKYILMAZ A., LIU Y., QIU X. & RADEV D. (2021). QMSum : A new benchmark for query-based multi-domain meeting summarization. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5905–5921, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.472.