DEBAT - DÉBAT - DEBATE

AI-rchivist : retour sur une expérience d'IA appliquée aux archives

Xavier Gillard

I. Introduction

Lorsqu'on parle d'histoire, et à fortiori lorsqu'on entame la lecture d'un article de la *Revue Belge d'Histoire Contemporaine,* il est fort probable qu'on ne s'attende pas à lire un texte écrit par un informaticien. Il n'aura pourtant échappé à personne que, depuis la présentation de ChatGPT au grand public par OpenAI fin 2022, il ne s'écoule plus un jour sans que nous soyons confrontés d'une façon ou d'une autre aux possibilités offertes par ces nouvelles technologies.[1] L'impact de celles-ci paraît tellement important que d'aucuns qualifient désormais l'intelligence artificielle (IA) de nouvelle révolution industrielle au même titre que la mécanisation et l'électrification.[2] Par ailleurs, une étude récente menée par le National Bureau of Economic Research indique que l'adoption de ces nouvelles technologies a été plus rapide auprès des individus qu'auprès des organisations ce qui semble indiquer que ces dernières cherchent toujours – dans leur grande majorité, à trouver une manière de bénéficier des gains de productivité offerts par ces nouveaux outils, tout en respectant les contraintes réglementaires et organisationnelles qui sont les leurs.[3] Le monde des archives n'échappe évidemment pas à cette tendance.

Face à la masse considérable de documents conservés et au défi humain, logistique et financier colossal que représente leur traitement, l'IA est souvent présentée comme une solution miracle. Cependant, l'enjeu n'est pas seulement technique. Si la numérisation massive est une première étape, elle ne résout pas le problème de l'accessibilité : les heuristiques de recherche complexes et l'impossibilité de "fouiller" le contenu textuel des images restent des obstacles majeurs pour le public et les chercheurs. Dès lors, la question centrale n'est pas de savoir si l'IA peut traiter des archives, mais comment l'intégrer de manière pertinente. Comment l'IA peut-elle devenir une alliée pour l'archiviste, un "assistant" capable de l'épauler dans ses tâches les plus chronophages sans pour autant effacer son expertise critique ?

Notre position dans ce débat n'est ni celle d'un techno-optimisme béat, ni celle d'un rejet de principe. Nous défendons l'idée selon laquelle l'IA peut légitimement trouver sa place au sein des institutions d'archives, à condition de ne pas masquer la complexité de sa mise en œuvre. Le déploiement d'outils réellement utiles aux archivistes et aux usagers ne s'improvise pas : il requiert d'abord des infrastructures de calcul coûteuses qui sont malheureusement encore souvent hors de portée des budgets culturels actuels. De plus, il exige ensuite une collaboration étroite où ingénieurs IA et archivistes co-construisent la solution. Surtout, ces technologies doivent être envisagées comme des leviers de productivité venant compléter l'expertise humaine, et non comme des substituts destinés à la remplacer. Enfin, cet enthousiasme prudent ne doit pas faire l'économie d'une vigilance face aux risques : l'IA n'est pas une panacée ; elle commet des erreurs, reproduit des biais statis-

1. OPENAI, Introducing ChatGPT, <https://openai.com/index/chatgpt/>, consulté le 9 avril 2025.
2. De la vapeur à l'IA : les trois révolutions industrielles qui ont façonné notre époque - RTBF Actus, site web de RTBF, <https://www.rtbf.be/article/de-la-vapeur-a-l-ia-les-trois-revolutions-industrielles-qui-ont-faconne-notre-epoque-11501099>, consulté le 22 mai 2025.
3. Bick, A., A. Blandin, & D. Deming, *The Rapid Adoption of Generative AI,* National Bureau of Economic Research, Cambridge (MA), 2024, 4-5, 29. DOI : https://doi.org/10.3386/w32966

tiques et peut induire, par son usage même, une confiance excessive chez l'utilisateur.

Ce document présente le résultat d'une expérience réalisée dans le cadre du projet ARKEY mené conjointement par les Archives de l'État en Belgique, et par l'UCLouvain.[4] Celle-ci a permis la création d'un outil nommé AI-rchivist qui vise à assister l'archiviste dans les tâches les plus ingrates de son travail. Après avoir détaillé les motivations du projet et son fonctionnement concret, nous aborderons les défis majeurs que cette approche soulève : ceux qui relèvent de la technique ou de la qualité des données, ceux qui sont inhérents à la langue, et ceux qui sont intrinsèquement liés à nos biais d'êtres humains.

II. Les motivations

Le développement d'outils basés sur l'IA tels que AI-rchivist que nous présenterons ci-dessous trouve sa justification dans la masse considérable de documents conservés au sein des fonds d'archives. La transcription, l'analyse, l'indexation et la mise à disposition de ces collections au public supposent de résoudre des défis énormes tant sur le plan humain qu'en matière d'infrastructure et de moyens financiers. À l'heure où la fréquentation des salles de lecture diminue, rendre ces pièces accessibles au public est devenu une priorité pour de nombreuses institutions en vertu de l'enjeu démocratique lié à la diffusion et à l'accessibilité de ce patrimoine commun.

Pourtant, dans un contexte de finances publiques limitées, les services d'archives sont rarement érigés en priorité, ce qui entraîne un manque de personnel qualifié, en particulier pour les tâches de paléographie et celles qui nécessitent une compréhension fine du contexte de production des documents. Face à ces contraintes, l'intelligence artificielle est envisagée comme une solution potentielle pour résoudre une partie de ces problèmes, et ce, à un coût potentiellement inférieur à celui du travail humain. Des projets de recherche récents tels que PARDONS, ACCESS, et ARKEY explorent activement cette voie, servant de terrain d'expérimentation pour le développement de nouveaux outils.

Le projet PARDONS vise à digitaliser, retranscrire et ouvrir à la recherche les vastes collections de lettres de grâce octroyées par les princes bourguignons et habsbourgeois. Mené par les Archives de l'État en collaboration avec la KU Leuven, l'UCLouvain et Histories vzw, ce projet analyse ces riches sources narratives pour offrir une histoire innovante du pouvoir et des relations sociales dans les anciens Pays-Bas.

De même, le projet ACCESS (Access to court files and access to justice), mené par les Archives de l'État et la VUB, se concentre sur les archives judiciaires de l'époque moderne (XV^{e}-XVIII^{e}). En s'appuyant notamment sur les fonds du Conseil de Brabant, il exploite ces procès et registres pour comprendre la vie quotidienne de la population, y compris les plus pauvres.

L'objectif du projet ARKEY est à la fois simple et ambitieux : il vise à utiliser des méthodes computationnelles (basées sur l'informatique) pour rendre les archives plus facilement accessibles au public et à la recherche. Bien qu'ils soient essentiels, les efforts de numérisation ne permettent pas à eux seuls de rencontrer cet objectif. En effet, la terminologie archivistique, les heuristiques de recherche, ainsi que le fait qu'il ne soit pas possible de chercher directement dans le contenu des documents numérisés sont autant d'écueils qui obstruent l'accès aux collections.

III. La vision

Il convient toutefois de souligner que, contrairement à une vision classique qui verrait le travail

4. ARKEY - AI meets archives - Archives de l'État en Belgique, <https://arch.arch.be/index.php?l=fr&m=nos-projets&r=projets-de-recherche&pr=arkey-ai-meets-archives>, consulté le 22 mai 2025.

humain être remplacé par des systèmes automatisés, notre approche place l'archiviste au centre du processus. L'objectif de cette approche – nommée *machine-in-the-loop* – n'est en aucun cas de substituer la machine à l'expertise humaine, mais bien de la renforcer grâce à un "assistant" numérique.[5] Cet outil est conçu pour multiplier les capacités de l'archiviste notamment en matière d'analyse et de traitement des sources historiques.

Pour donner corps à cette vision, un outil – nommé AI-rchivist – a été développé en collaboration avec les chercheurs travaillant sur les projets PARDONS et ACCESS.[6] [7] L'objectif de AI-rchivist, est d'utiliser un modèle d'IA générative afin de faciliter l'extraction de métadonnées, telles que les lieux, personnes et dates, mentionnées dans les textes originaux, et de rédiger un résumé succinct de chaque document soumis à AI-rchivist dans les trois langues nationales ainsi qu'en anglais. Les données ainsi obtenues sont stockées dans une base de données relationnelle qui doit en faciliter l'export et l'intégration avec d'autres systèmes informatiques.

Dans cet outil, l'approche "machine-in-the-loop" adoptée se concrétise par deux aspects essentiels : si c'est bien la machine qui se charge de faire une première proposition d'extraction de données, les contenus d'absolument tous les champs présents dans l'interface fournie à l'utilisateur sont éditables.[8] Cela signifie que l'archiviste peut compléter, vérifier, et le cas échéant corriger les erreurs qu'il observe avec très peu d'efforts. Par ailleurs, la plateforme incorpore un chat-bot au travers duquel l'archiviste peut faire part de ses remarques à l'outil, ce qui lui permet par exemple, de demander à l'IA d'apporter des corrections transversales aux données produites. Pour traiter les cas les plus extrêmes, l'outil incorpore une fonctionnalité permettant à l'archiviste de remplacer le modèle d'intelligence artificielle (LLM) par un autre plus puissant (mais plus lent et/ou plus cher à utiliser).

IV. Corpus

Pour illustrer concrètement l'application et les défis liés à la mise en œuvre de l'outil AI-rchivist, une étude de cas a été menée sur un corpus de 419 textes. Comme indiqué par la Figure 1, ce corpus comporte une très large majorité de lettres de pardon parmi lesquelles seule une petite fraction (environ 10% des textes) est rédigée en français moyen. Le reste étant rédigé en néerlandais moyen. Il est toutefois important de noter que cette catégorisation par langue n'est pas parfaitement stricte, puisqu'il est fréquent que des phrases ou notes rédigées dans l'autre langue soient insérées dans le texte – ce qui est généralement perçu comme une complication supplémentaire pour le développement de systèmes de traitement de la langue naturelle.

Intégrer toutes les étapes du traitement IA des archives, de la numérisation jusqu'à l'extraction des métadonnées dans une seule plateforme aurait été idéal. Toutefois, bien que des modèles d'IA soient mis au point via la plateforme Transkribus dans le cadre des projets PARDONS et ACCESS, ceux-ci n'étaient pas encore considérés comme étant suffisamment aboutis pour fournir des transcriptions fiables.[9] C'est pourquoi il a été décidé d'utiliser les transcriptions faites par des bénévoles ayant contribué à ces projets.

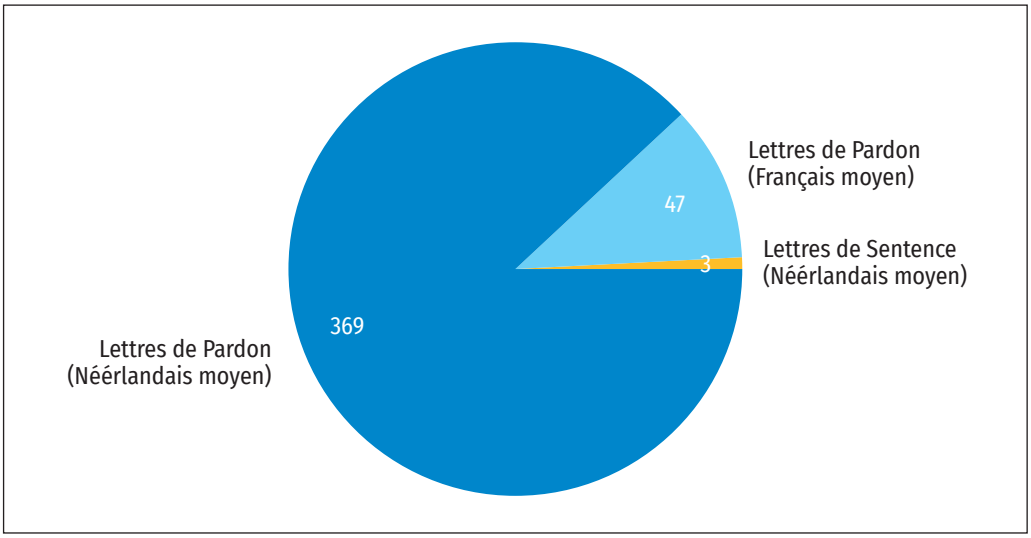
5. Clark, E. et al., "Creative Writing with a Machine in the Loop : Case Studies on Slogans and Stories" in *23rd International Conference on Intelligent User Interfaces,* ACM (Tokyo Japan), 2018, 329330. DOI : https://doi.org/10.1145/3172944.3172983. URL : <https://dl.acm.org/doi/10.1145/3172944.3172983>

6. PARDONS, <https://pardons.eu/fr/>, consulté le 17 juin 2025.

7. ACCESS to court files and access to justice. The Council of Brabant during the early modern era - Rijksarchief in België, <https://www.arch.be/index.php?l=nl&m=lopend-onderzoek&r=onderzoeksprojecten&pr=access-to-court-files-and-access-to-justice.-the-council-of-brabant-during-the-early-modern-era>, consulté le 17 juin 2025.

8. Clark, E. et al., "Creative Writing with a Machine in the Loop", 329331.

9. Kahle, P. et al., "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR),* IEEE Computer Society (Kyoto, Japan), 2017. DOI : https://doi.org/10.1109/ICDAR.2017.307. URL : <https://ieeexplore.ieee.org/abstract/document/8270253>.



Description du corpus de l'étude de cas.

Les textes de ce corpus sont donc le résultat du travail minutieux de volontaires qualifiés et expérimentés, mais ils ne sont évidemment pas "parfaits". En effet, comme on peut légitimement s'y attendre, certaines parties de ces textes sont illisibles ou rendues ambiguës par l'absence de ponctuation et l'utilisation non-standardisée de la casse (majuscule/minuscule) ce qui en complique l'interprétation.

Ce choix de privilégier les transcriptions humaines a été motivé par la volonté de permettre aux trois projets (ARKEY, PARDONS et ACCESS) d'avancer de concert, sans être freinés par les limitations techniques des uns et des autres. L'étude de cas vise donc à évaluer la capacité de AI-rchivist à extraire des métadonnées pertinentes à partir de textes complexes tout en mettant en lumière les défis spécifiques liés au traitement de langues anciennes et aux contraintes liées à la qualité des données.

V. AI-rchivist

Afin de permettre au lecteur d'avoir une meilleure idée des possibilités offertes par notre prototype, cette section reproduit une série de captures d'écran issues de l'application elle-même. Afin d'en améliorer la compréhension, toutes ces images utilisent le même exemple tiré du corpus. En l'occurrence, il s'agit de la lettre de rémission par laquelle l'Empereur Charles Quint a accordé son pardon à Huguenin Moreau en juin 1521.[10] Cet exemple illustre bien la complexité du corpus avec lequel nous avons travaillé. En effet, outre le fait que le texte soit rédigé en utilisant la langue et le style des documents juridiques du XVI^{e} s., celui-ci mélange aussi largement les passages rédigés en français et en néerlandais – ce qui en complexifie la lecture même pour un humain qui serait familier avec les versions actuelles de ces deux langues.

La Figure 2 ci-après montre l'écran d'accueil de la plateforme. On peut y observer que l'interface proposée est somme toute assez simple : une boîte de texte dans laquelle l'archiviste peut coller la transcription du document sur lequel il travaille, un bouton pour lancer l'analyse du texte par l'IA ainsi qu'un bouton permettant de faire apparaître la boîte de dialogue permettant d'interagir avec le système.

10. Huguenin Moreau (juni 1521) – PARDONS, <https://pardons.eu/nl/2023/06/letter-of-grace-for-pour-huguenin-moreau-june-1521/>, consulté le 20 juin 2025.

Les Figure 3, Figure 4, et Figure 5 montrent comment se présentent les autres onglets de l'application. Sur la Figure 3, on peut voir que, malgré les difficultés inhérentes à ce texte, le système a été capable de retrouver automatiquement une série d'informations pertinentes telles que le type de document ou la date des faits. On y trouve aussi un bref résumé du contenu du texte dans les trois langues nationales et l'anglais. Les Figure 4, et Figure 5 sont d'aspect fort similaire et permettent de se faire rapidement une idée des personnes et lieux mentionnés dans le texte original ainsi que de leur rôle.

Il est important de rappeler que l'approche choisie maintient l'archiviste dans son rôle de garant de la pertinence et de la validité des informations reprises et qu'à cet effet, tous les champs visibles sont éditables. Il est possible de corriger aussi bien le résumé du texte que d'ajouter, modifier ou supprimer des personnes et des lieux.

La Figure 6 nous donne un exemple d'interaction possible avec AI-rchivist. Par ce biais, il est possible d'apporter des modifications transversales à toutes les données extraites par le système.

VI. Approche et limitations

La démonstration donnée ci-dessus pourrait sembler "trop belle pour être vraie", et bien que le prototype soit fonctionnel, il est essentiel d'en comprendre le fonctionnement pour mieux en cerner les forces et les faiblesses.

L'approche utilisée par AI-rchivist s'appuie essentiellement sur les capacités langagières d'un gros modèle de langage (LLM) pour comprendre à la fois le texte ancien et suivre les instructions qui lui sont données pour effectuer l'extraction des métadonnées à proprement parler. Cette approche est simple, rapide à mettre en œuvre et elle peut s'avérer utile pour soutenir les archivistes dans leurs tâches les plus chronophages. Elle se heurte néanmoins à des limitations de quatre ordres :

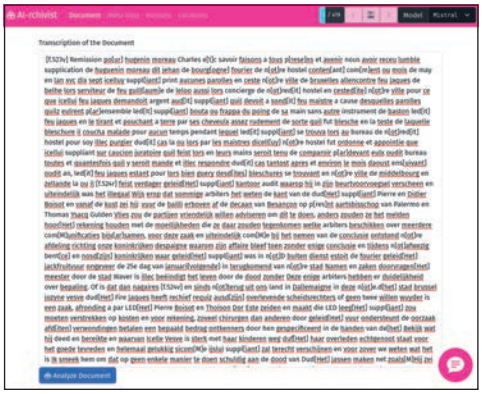
1. La puissance de calcul nécessaire
2. Les hallucinations[11]
3. L'utilisation de langues anciennes
4. Le facteur humain
5. La perte de contexte

La puissance de calcul nécessaire

L'un des défis majeurs liés à l'utilisation des grands modèles de langage (LLM) réside dans la puissance de calcul colossale qu'ils requièrent. En effet, ces modèles comportent plusieurs dizaines ou centaines de milliards de paramètres. Ceux-ci correspondent aux connaissances qui ont été « apprises » par le modèle lors de son entrainement. Il s'agit de nombres réels qui sont agencés en une série de matrices et qui permettent au modèle de produire un résultat utile. Le résultat est obtenu en combinant lesdites matrices avec le texte qui lui est donné en entrée via des opérations mathématiques – essentiellement des multiplications.[12] [13]

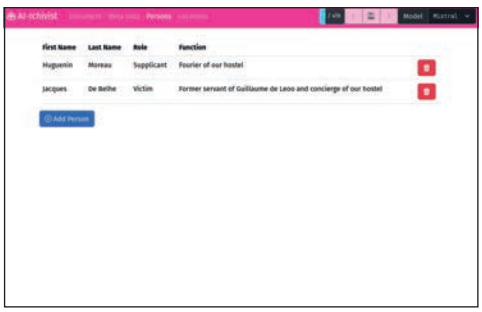
Le lecteur ayant un souvenir de ses cours de calcul matriciel à la fin du secondaire aura alors remarqué que ces opérations nécessitent de calculer un nombre phénoménal de multiplications d'additions qui sont toutes assez simples. Celles-ci sont par ailleurs largement indépendantes les unes des autres, ce qui signifie que ces calculs très intensifs peuvent être accélérés en réalisant toutes ces opérations mathématiques en parallèle les unes des autres (en même temps) plutôt que séquentiellement (l'une après l'autre). C'est pourquoi l'intelligence artificielle utilise des puces dédiées appelées GPU dont il est aujourd'hui impossible de se passer.[14]

11. Xu, Z., S. Jain, & M. Kankanhalli, *Hallucination is Inevitable : An Innate Limitation of Large Language Models*, arXiv, 2025. DOI : https://doi.org/10.48550/arXiv.2401.11817.
12. Du point de vue de la terminologie, il serait plus exact de parler de tenseurs puisqu'il s'agit souvent de structures ayant plus deux dimensions.
13. Ou l'image, le son, etc.
14. Graphics Processing Unit car ces puces ont initialement été développées pour les jeux vidéo.

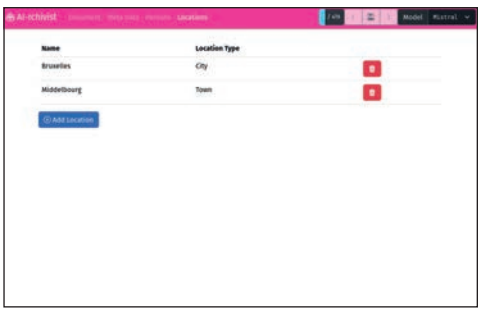


L'écran d'accueil permet de traiter le document en y copiant simplement la transcription du texte.

L'onglet des métadonnées qui donne un bref aperçu de la pièce traitée.



L'onglet des personnes reprend le détail des personnes identifiées.



L'onglet des lieux donne un aperçu des lieux mentionnés dans le texte.

Exemple d'interaction visant à apporter des corrections transversales.

L'analogie suivante est utile pour comprendre la différence qui existe entre les GPU et les CPU qui équipent nos ordinateurs et téléphones portables.[15] Les CPU peuvent être comparés à des bateaux hors-bords : ils sont très rapides et très maniables mais ils n'offrent que peu de place pour les personnes et le stockage. Les GPU, à l'inverse, peuvent être comparés à des bateaux cargo. Ils sont généralement plus lents, moins maniables, consomment plus d'énergie mais ils offrent une capacité de transport inégalée. En suivant cette analogie, on doit alors comparer les calculs à réaliser par le modèle d'IA à des conteneurs qui doivent être transportés d'un point à un autre. Ce qui rend la nécessité de recourir à ces puces spécialisées encore plus évidente.

Outre le rôle de chacun de ces composants, l'analogie ci-dessus permet de mettre en lumière un autre aspect qui différencie les deux types de puces : leur coût. Là où le prix des meilleurs CPU plafonne aux alentours d'une dizaine de milliers d'euros, le prix unitaire des GPU destinés au marché professionnel varie entre 15 000 et plus de 30 000 euros.[16] Dans la plupart des cas, le déploiement d'un grand modèle de langage requiert que plusieurs de ces GPU haut de gamme soient utilisés ensemble.

Le coût d'acquisition, la consommation énergétique et le personnel nécessaire à la maintenance de tels systèmes représentent une charge financière considérable pour les organisations qui souhaiteraient s'équiper pour opérer elles-mêmes de grands modèles de langages. C'est pourquoi l'exploitation des grands modèles de pointe est désormais l'apanage d'un nombre limité d'acteurs mondiaux qui louent leurs services au travers de solutions cloud.

Il est néanmoins possible de déployer de plus petites versions de ces grands LLM sur du matériel moins cher (de l'ordre de quelques milliers d'euros) en acceptant une certaine dégradation des performances et capacités de l'IA déployée. C'est cette approche qui a été choisie dans le cadre du développement de AI-rchivist qui utilise des "petits" LLM – parfois appelés SLM pour *Small Language Model,* ne comprenant « que » quelques milliards de paramètres.

Les hallucinations

Les hallucinations sont l'un des défis majeurs auxquels font face les utilisateurs d'applications basées sur les LLM. Il s'agit du problème qui survient lorsque le modèle *génère du texte qui semble plausible mais qui est factuellement incorrect ou insensé.*[17] Ce problème est bien connu et il affecte plus fréquemment les petits modèles que les grands.

Il est communément admis que l'utilisation de techniques qui demandent au modèle d'extraire de l'information parmi un contexte fourni permettent de réduire le risque d'hallucinations. Mais il n'en demeure pas moins qu'il a récemment été prouvé que ce phénomène était inévitable.[18] Par ailleurs, il faut aussi considérer les omissions qui pourraient être le fait du système d'IA et qui pourraient invisibiliser certaines personnes, faits ou lieux. Ces deux aspects sont particulièrement problématiques lorsqu'ils surviennent dans un cadre archivistique puisque l'un des rôles essentiels de cette discipline est justement de garantir la qualité des données.

L'évaluation à postériori des données extraites par AI-rchivist sur le corpus de texte a révélé que les hallucinations et les omissions sont un problème important dans le cadre considéré. La plupart des critiques formulées par les archivistes qui ont accepté de participer à cette évaluation concer-

15. Central Processing Unit, l'unité centrale de l'ordinateur.
16. Certaines cartes graphiques à destination du grand public comme par ex. les NVIDIA RTX 3090, 4090, 5080 et 5090 peuvent aussi être utilisées à cet effet, ce qui fait qu'elles sont désormais assez rares et chères.
17. Xu, Jain & Kankanhalli, *Hallucination is Inevitable,* 13.
18. *Idem,* 2,7 & 9-10.

naient le fait que, sans être tout à fait incorrectes, les données fournies par l'outil étaient perfectibles puisqu'elles mentionnaient des éléments qui n'avaient pas lieu d'être.

L'utilisation de langues anciennes

Le fait de travailler avec des langues anciennes est également une source de difficultés pour les systèmes d'IA. De la même façon qu'un être humain du XXIe s. – même spécialiste – sera moins familier avec les langues anciennes qu'avec les langues modernes, les systèmes d'IA ont essentiellement été entrainés en utilisant d'énormes jeux de données récentes tels que le *Common Crawl Corpus* qui reprend le contenu de plus de 250 milliards de pages web.[19] De ce fait, l'exposition des grands modèles de langages aux langues anciennes est relativement faible.

Comme nous l'avons vu lors de la démonstration présentée dans les sections précédentes, cette faible exposition n'empêche pas que l'utilisation de LLM puisse s'avérer utile pour traiter des textes très anciens. Toutefois, la performance qu'on peut attendre de ces modèles se trouve réduite pour deux raisons. La première – et sans doute la plus intuitive à comprendre – est qu'étant donné le faible volume de textes anciens que ces modèles ont pu rencontrer lors de leur phase d'entraînement, la finesse avec laquelle ils ont pu intégrer les nuances de la langue utilisée dans ces textes est elle aussi amoindrie.

Une brève introduction au fonctionnement de l'IA

La seconde raison est plus technique et impose une véritable limite à ce qui peut être envisagé en termes d'utilisation de LLM. Pour bien l'appréhender, il est nécessaire de comprendre – même à un assez haut niveau – comment fonctionnent ces modèles de langages. A cet effet, la première chose qu'il est important de savoir est que, contrairement à nous, les machines n'ont aucune notion de ce qu'est un mot. Elles ne savent manipuler qu'un ensemble fini de nombres.

Pour pallier cette incompatibilité fondamentale, les systèmes informatiques utilisent toujours un vocabulaire de taille fixe. Il s'agit d'une grande table de correspondance qui permet d'identifier chaque mot par un nombre unique. Mistral, par exemple, fonctionne avec un vocabulaire de 32 000 mots. Parmi ceux-ci, le mot 'histoire' correspond au nombre 25 291 et le mot 'archive' correspond au nombre 23 682.[20]

La valeur des nombres associés à chaque mot n'a absolument aucune importance. L'unique rôle de ces valeurs est d'identifier chacun des mots du vocabulaire afin de permettre à l'ordinateur de passer d'un monde de mots – qu'il est incapable d'exploiter – à un monde de nombres qui sont plus faciles à manipuler.

Cependant, le vocabulaire utilisé par un modèle de langage ne pourrait se limiter à ce qui est défini dans le dictionnaire de référence. En effet, pour pouvoir générer du texte qui ressemble à celui qu'un être humain aurait pu écrire, les LLM doivent pouvoir tenir compte des inflexions grammaticales et de la puissance créatrice des langues vivantes qui combinent et inventent constamment de nouveaux mots. C'est pourquoi un « mot » du vocabulaire utilisé par les modèles de langage ne correspond pas exactement à un mot tel que nous l'entendons. Il s'agit plutôt de sous-mots qui pourraient être porteurs de sens. C'est pour cette raison qu'on utilise le terme « token » plutôt que « mot » lorsqu'on veut désigner un élément du vocabulaire.

La toute première étape du traitement réalisé par un LLM consiste donc à découper le texte qu'on lui soumet (aussi appelé le prompt) en une suite de tokens appartenant à son vocabulaire. A titre d'exemple, Mistral découpe la phrase « j'essaie de ne pas vous perdre avec la technique » comme ceci :

19. Common Crawl - Open Repository of Web Crawl Data, <https://commoncrawl.org>, consulté le 26 juin 2025.
20. Jiang, A. Q. et al., *Mistral 7B*, arXiv, 2023. DOI : https://doi.org/10.48550/arXiv.2310.06825.

j	‘	essa	ie	de	ne	pas	vous	perd	re	avec	la	technique

Exemple 1 : Découpe de la phrase “J’essaie de ne pas vous perdre avec la technique” par Mistral

Il est intéressant de noter dans cet exemple que la ponctuation est traitée comme un “mot” du vocabulaire. Il est aussi intéressant d’observer que la découpe n’est pas très éloignée de celle qu’un humain aurait pu réaliser moyennant le fait que le radical et la terminaison des verbes ait été séparée.

Le second élément qu’il convient de comprendre pour appréhender les limites “dures” liées à l’utilisation des LLM tient à la façon dont ceux-ci produisent leurs réponses. Contrairement à nous, les modèles ne commencent pas par « comprendre » le texte (prompt) qui leur est donné pour ensuite rédiger leur réponse. Ils essaient simplement de continuer le prompt en y ajoutant chaque fois un seul mot.[21] Ce qui signifie que pour produire une seule phrase de texte, une IA devra relire plusieurs fois tout le texte qui lui a été donné ainsi que l’ensemble du texte qu’elle a déjà généré. Cette façon de fonctionner est illustrée par la Figure 7.

Bien que les LLM fassent preuve d’une capacité conversationnelle impressionnante, il y a une limite à la taille du texte avec lequel ils peuvent travailler. Cette limite s’appelle la *taille de contexte :* il s’agit du nombre de « mots » que le modèle est capable de garder en mémoire lorsqu’il cherche à trouver la meilleure façon de continuer le texte. Une fois que cette limite est atteinte, le modèle devient incapable d’écrire quoi que ce soit. Et de ce fait, il sera tout simplement incapable de commencer à répondre à la question qui lui est posée.

En quoi est-ce problématique ?

Comme cela a été exposé précédemment, les langues anciennes sont assez peu présentes dans les corpus de textes utilisés pour entraîner les modèles de langage. De ce fait, les variations orthographiques qui sont typiques de ces textes anciens n’ont pas pu être prises en compte lors de la création du vocabulaire du LLM. Il en résulte donc que la découpe obtenue pour ces textes sera moins bonne et découpera les mots en plus petits morceaux. Dans les cas extrêmes, on peut même imaginer que le texte soit découpé lettre par lettre.

Nos expériences montrent que les textes anciens consomment environ 2 fois plus de « mémoire » (tokens) que les textes modernes à cause de leur vocabulaire spécifique et de leur trop grande variation orthographique. Par conséquent, la probabilité que la réponse apportée par le modèle à propos de ces textes anciens soit une hallucination s’en trouve renforcée. Par ailleurs, nos expériences ont aussi révélé que tous les LLM publiés avant août 2024 ne disposaient pas d’assez de mémoire pour lire la totalité des documents avant de commencer leur analyse.

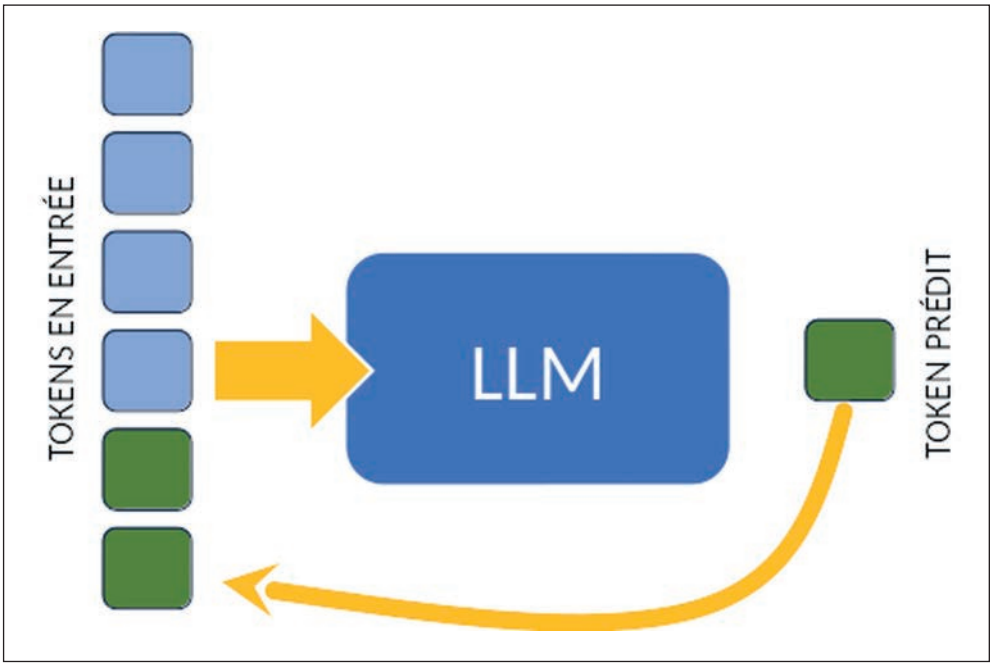
Certaines techniques telles que la compression de prompt dont l’objectif est de supprimer automatiquement un certain nombre de mots « inutiles » ont été utilisées pour tenter de remédier à ce problème.[22] Elles n’ont toutefois eu qu’un effet assez marginal sur la quantité de textes pouvant réellement être traités tout en provoquant une dégradation sensible de la qualité des données extraites.

Le facteur humain

Le facteur humain peut lui aussi s’avérer problématique lors de l’implémentation d’une technologie telle que AI-rchivist. Les études récentes ont mis en évidence que l’introduction de systèmes d’IA pouvait amener une modification de la qua-

21. Token si on veut être précis.

22. Pan, Z., Lun-Wei, K., A. Martins, & V. Srikumar (Éd.), *LLMLingua-2 : Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression* in Findings of the Association for Computational Linguistics : ACL 2024, Association for Computational Linguistics, Bangkok, 2024. DOI : https://doi.org/10.18653/v1/2024.findings-acl.57. URL : <https://aclanthology.org/2024.findings-acl.57/>, pages 967968.



Génération de texte par un LLM.

lité des pratiques professionnelles en raison du *biais d'automatisation*.[23] Ce biais est causé par le fait que les êtres humains font généralement trop confiance aux systèmes d'IA, ce qui les pousse à s'aligner sur les résultats suggérés par ceux-ci – même lorsque ces résultats entrent en conflit avec leurs propres conclusions.

Cette confiance disproportionnée envers les systèmes d'IA peut s'avérer spécialement problématique lorsqu'on l'applique au traitement de sources archivistiques. En effet, le rôle de l'archiviste est de sélectionner, conserver et rendre accessibles ces documents aux générations futures. Et, bien que les capacités des systèmes d'IA actuels soient impressionnantes, nous avons montré précédemment qu'elles ne sont pas exemptes de défauts et peuvent donner des résultats erronés. S'aligner sur les prédictions faites par l'IA dans ces cas-là provoquerait, au mieux, une dégradation de la qualité des données et des erreurs d'indexation. Mais les conséquences pourraient être irrémédiables si elles se produisent dans le contexte de la procédure visant à sélectionner les sources qui doivent être conservées à titre d'archives permanentes et celles qui peuvent être détruites.

Outre le biais d'automatisation, le facteur humain peut aussi s'avérer être un coût significatif lors de l'introduction de tels systèmes qui automatisent le traitement des archives. En effet, en consacrant une partie significative de son temps à lire, classer, et indexer des fonds d'archives, un archiviste s'approprie en profondeur le contenu de ce fonds. Et c'est cette appropriation qui peut être mise au service du public et de la recherche pour tisser des liens entre différents fonds, ou pour retrouver certaines pièces peu connues. Une trop grande délégation du traitement de la masse des documents à des machines pourrait donc engendrer une perte au niveau de l'appropriation de ces corpus par les archivistes. Ce qui pourrait représenter un coût non seulement pour la profession qui serait moins confrontée au matériau historique, mais aussi un coût démocratique pour la société au sens plus large.

La perte de contexte

La dernière limite qu'il semble important de mentionner est d'ordre méthodologique. En effet, si les méthodes computationnelles permettent d'extraire les métadonnées relatives à chaque document avec une finesse inégalée, elles se heurtent frontalement à la pratique archivistique moderne. Les algorithmes ne travaillent jamais qu'au niveau d'une seule pièce à la fois, alors que les archivistes s'attellent à faire émerger le *contexte* de création d'un fonds d'archives. C'est ce contexte qui « colore » le sens à attribuer à chaque document présent au sein d'un même fonds.[24] C'est aussi ce *contexte* qui, une fois cristallisé dans les guides puis les jalons de recherche, sert de base heuristique à la recherche puis à l'exploration des fonds en vue de répondre à une question précise. En perdant ce contexte au profit d'une analyse automatique – même très fine – de chacun des documents, il est possible que nous créions plus de données (brutes et non interprétées) que d'information (par essence contextualisée).

La question de savoir comment approcher cette question n'a – à la connaissance de l'auteur, pas encore été examinée. Elle était par ailleurs au cœur d'un projet ayant été soumis à financement cet hiver.

VII. Conclusions

Cet article a commencé par présenter AI-rchivist, un outil innovant destiné à soutenir le travail archivistique dans ses tâches les plus chronophages. Nous avons montré qu'il permettait de faciliter ce travail même en étant confronté à des textes rédigés en français et néerlandais moyens (combinés). Nous avons aussi mis en lumière les limitations

23. Horowitz, M. C. & L. Kahn, "Bending the Automation Bias Curve: A Study of Human and AI-based Decision Making in National Security Contexts", *International Studies Quarterly*, 2/68, 2024.
24. Il n'est pas impossible que plusieurs copies d'un même document soient présentes au sein de différents fonds. C'est toutefois, le contexte de création de ce fonds qui affinera la compréhension qu'il faudra avoir de ce document au sein du fonds.

intrinsèques liées au déploiement d'outils comme AI-rchivist. A savoir : le coût significatif de ces systèmes imposé par la puissance de calcul qu'ils requièrent, le phénomène des hallucinations qui peuvent amener à des analyses factuellement erronées bien que vraisemblables. Mais aussi, les limites – y compris techniques – de l'utilisation de grands modèles de langages lorsqu'ils sont utilisés pour traiter des textes anciens. Nous avons aussi montré que le facteur humain pouvait être à la fois un risque et un coût substantiel lors de l'introduction de systèmes d'IA.

Ces considérations sortent complètement du cadre de l'ingénierie qui rend possible la réalisation de tels systèmes d'IA. Elles interrogent aussi en profondeur le rôle de l'archiviste et l'approche traditionnelle de l'archivistique qui met l'accent sur la contextualisation des pièces. En travaillant au niveau de chaque pièce, il est possible de réaliser une extraction de métadonnées d'une finesse jamais égalée. Mais aussi utiles que soient ces métadonnées très fines, l'approche basée sur chaque pièce plutôt que sur les fonds pourrait rendre plus difficile la recherche et la création de liens entre différents fonds.

Xavier Gillard est chercheur FED-tWIN aux Archives de l'État et à l'UCLouvain. Email : Xavier.Gillard@arch.be.