Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Archives as wellspring, verifier and generator of knowledge: a vision for the age of AI

Woojin Jung

presolving@naver.com

Yulgok Institute of Korean Studies

Ilhoon Oh

Yulgok Institute of Korean Studies

Chunghyun Kim

Yulgok Institute of Korean Studies

Jongrok Lee

Yulgok Institute of Korean Studies

Joonmo Ahn

Narainformation Co. Ltd.

Gyu-min Lee

Narainformation Co. Ltd.

Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11509514/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Archives and cultural heritage institutions increasingly offer services that combine large language models with retrieval-augmented generation, allowing users to put questions to their collections and receive answers grounded in archival records. Existing research has treated AI either as a tool for archival processes or as a source of knowledge production to be restrained, leaving open what archives should aim to become in the age of AI. Archival science, however, has already opened the way towards knowledge generation, at least at the level of theory, by regarding the archive not as a mere repository of records but as a site where the meanings of records are continually reconstructed. This paper proposes the knowledge-generating archive, which generates knowledge in response to users' questions on the basis of its records, and explores the role of archives in the age of AI more actively by examining its strengths and limitations through a case study of letters from Korea's Joseon dynasty. It proposes three principles for the knowledge-generating archive: generation and verification, interdependent internal principles of the archive, and the wellspring, an external principle through which the archive supplies human-made sources of assured provenance to a knowledge ecosystem increasingly filled with AI-generated text.

Introduction

Major archival and cultural heritage institutions around the world have been launching new services that incorporate generative artificial intelligence (AI). The National Library of Luxembourg has introduced conversational search for its digitised collection of historical newspapers (Luxembourg Government, 2023), and the Historical Archives of the European Parliament have released Archibot, a service based on a large language model (LLM) that allows users to ask questions about parliamentary documents dating back to 1952 in multiple languages and to receive answers together with the source documents (Davies, 2024). The US National Archives and Records Administration (NARA) has also pursued the introduction of a public-facing chatbot, Archie AI (Koebler, 2024). Finn and Khosrowi (2026) recently offered a critical examination of the significance and problems of such services. As LLMs reshape how information is gathered, accessed and shared online (Burton et al., 2024), the knowledge ecosystem is expected to undergo profound change, and critical scrutiny of the kind Finn and Khosrowi provide is necessary. Critique alone, however, cannot tell us what archives should aim to become in the age of AI. This paper proposes a vision for archives in the age of AI.

The technical foundation of Archibot and the other services mentioned above is retrieval-augmented generation (RAG), formalised by Lewis et al. (2020). Lewis and his colleagues combined the 'parametric memory' that a language model stores within its parameters with a 'non-parametric memory' that exists as external documents and is accessed through retrieval. In their experiments, the external memory was a vector index of Wikipedia articles, which offered two advantages: the grounds of a response could be checked, and knowledge could be updated simply by replacing the index. Vector retrieval, however, is weak at following relationships between entities and at answering questions directed at a collection as a whole, such as what its main themes are. To address this, graph-based RAG has emerged, which

organises entities and their relationships into a knowledge graph and uses it for retrieval (Edge et al., 2024). In this paper, we call an archive implemented through such a combination of technologies, namely vector embeddings that find semantically related records, a knowledge graph that links entities, and an LLM that generates responses from the evidence retrieved from both, a 'knowledge-generating archive'. A knowledge-generating archive is an archive that generates knowledge in response to users' questions on the basis of the records it holds, and that is equipped with a system for verifying the grounds of that knowledge.

RAG, which forms the foundation of the knowledge-generating archive, was not originally designed for archives. It began as an attempt to reduce hallucination in LLMs by having them retrieve trustworthy external sources before generating a response and ground their responses in those sources (Lewis et al., 2020). Hallucination is an inherent limitation of LLMs (Kalai et al., 2025), and RAG is, in effect, a remedy devised to address this limitation. Yet RAG, once applied to archives, is experienced as the archive itself. The European Parliament released Archibot under the name 'Ask the EP Archives' (Davies, 2024), and, as Finn and Khosrowi point out, users take it as a direct conversation with the archive (Finn & Khosrowi, 2026). Indeed, Finn and Khosrowi's critique itself rests on the premise that these tools pose problems not of AI but of archives. Ordinary users and experts alike, it seems, understand such services within the category of the archive.

If such services are already experienced as part of the archive, their role can be examined from the perspective of the archive. An archive combined with LLMs through RAG can, and should, become a producer of knowledge that demands a high degree of trust, such as scholarly knowledge. Moreover, as AI-generated text increasingly fills the knowledge ecosystem, the value of archives holding human-made source records of assured provenance grows rather than diminishes. Research showing that AI models degenerate when repeatedly trained on data they themselves have generated supports this (Shumailov et al., 2024). From this perspective, this paper asks what vision archives should hold in the age of AI, and seeks the answer in three principles: generation and verification, interdependent internal principles through which the archive constructs knowledge and confirms its grounds, and the wellspring, an external principle through which the archive supplies human-made sources to the knowledge ecosystem. Of these, the wellspring and verification renew values long upheld by archives, while generation is a new principle demanded by the age of AI. To this end, the paper combines a conceptual analysis, drawing on archival theory, Aleida Assmann's theory of cultural memory and recent findings in AI research, with a case study of letters (ganchal) from Korea's Joseon dynasty (1392–1910).

The paper is structured as follows. The literature review examines existing research on archives and AI, showing that it has treated AI either as a tool for archival processes or as a source of knowledge production to be restrained, while leaving open the question of what archives should aim to become in the age of AI. The following section argues that the knowledge-generating archive realises in practice the paradigm shift already made in archival theory, and draws on Assmann's distinction between storage and functional memory to explain its dual character of preservation and generation. A case study then compares an LLM trained on letters received by a single individual in the Joseon period with a

knowledge-generating archive holding the same records, demonstrating the differences between them. Building on this, the paper presents a vision for archives in the age of AI through three principles, the wellspring, verification and generation, and shows how the two internal principles of generation and verification sustain each other within the archive while the external principle of the wellspring sustains the knowledge ecosystem. The conclusion summarises the contributions and limitations of the paper and outlines directions for future research.

Literature review

Research on issues arising at the intersection of archives and AI has grown rapidly in recent years. Although these studies often address several themes at once and are therefore difficult to divide strictly, they can be grouped into three strands according to their central focus. The first comprises function-oriented studies that treat AI as a tool for solving the challenges of records management. The second comprises value-oriented studies that ask which values must be upheld when archives and AI meet, or argue that archival principles should be applied to the data on which AI is trained. The third, which has emerged only recently, addresses the phenomenon of archives becoming directly involved in knowledge production through AI.

Most function-oriented studies treat AI as a tool for addressing long-standing challenges in records management. Colavizza et al. (2021), for example, surveyed efforts to automate the classification, appraisal and selection of records, the identification of sensitive information and the generation of metadata, setting out the changes and debates that AI may bring to records management as a whole. Shinde et al. (2024) showed that most existing studies have concentrated on tasks such as records classification and metadata processing. The volume edited by Jaillant (2022) brings together research in this direction, covering topics such as the automatic classification of images in photographic archives, the opening of closed born-digital records and the use of handwritten text recognition (HTR). The InterPARES project, which has studied the long-term preservation of authentic electronic records since 1999 (Duranti, 2005), is exploring in its latest phase, InterPARES Trust AI, how AI can be used to address challenges in records management and archives (Frontoni et al., 2022). Discussions of how AI can automate archival processes in this way lead in turn to questions of value.

Jaillant and Rees (2022) examined the problems of accountability, control and bias that may arise when AI is applied to archives, together with ways of overcoming them, while Colavizza and Jaillant (2026) argue that archival concepts such as authenticity and reliability should guide how AI is applied to archives. Whereas these studies ask about the values at stake in applying AI to archives, other studies work in the opposite direction, drawing on archival principles to reflect critically on the data and models of AI. Jo and Gebru (2020) argued, in pioneering work, that the approaches archives have long developed with regard to consent, power, inclusivity, transparency, and ethics and privacy could help mitigate bias and opacity in the construction of datasets. Desai et al. (2024) treated pretraining datasets collected from the web as ‘informal archives’ that mediate access to knowledge, and showed that the technical work of filtering data in fact involves value judgements, much like archival appraisal and selection. More

recently, attempts have emerged to compare the archival concepts of authenticity, reliability, integrity and usability with frameworks for ‘trustworthy AI’, redefining the trustworthiness of AI not as a merely technical property but as a matter of data governance grounded in provenance and lifecycle accountability (Barbudo Carrasco et al., 2026).

The studies discussed above can be said to show a healthy relationship between archives and AI, but they do not address the knowledge-generating dimension that archives acquire through their combination with AI. Research that treats archives as major actors in the knowledge ecosystem, that is, research addressing the phenomenon of archives becoming directly involved in knowledge production through generative AI, has emerged only recently. Finn and Khosrowi (2026) analysed RAG-based AI assistants in archives and argued that these tools disrupt the concepts surrounding the production of historical knowledge. Framing the user’s experience as a direct conversation with the archive, they contend, obscures the mediated nature of sources and the incompleteness of the archive, and leads users to bypass the processes of source criticism and contextualisation that historical research requires. Calling this the lure of instant history, they propose that the outputs of archival AI be described not as historical knowledge but as document syntheses, and that their limitations be clearly disclosed(Finn and Khosrowi, 2026, p. 10–12). Their study is significant in that it squarely acknowledges that AI-based archives have become actors in knowledge production. Its conclusion, however, leans towards guarding against the knowledge-producing function of archives.

Function-oriented studies have addressed the automation of archival processes and improvements in user access, while value-oriented studies have addressed ways of overcoming the problems that arise when AI is applied to archives and the application of archival principles to the problems of AI itself. Finn and Khosrowi (2026), by contrast, focus on the knowledge-generating dimension, but stop at pointing out its side effects and proposing norms to restrain them. This paper likewise focuses on archives as producers of knowledge, but takes a different direction. It starts from the premises that generative AI has already become a major actor in knowledge production, that archives will inevitably operate in relation to AI, and that AI-based knowledge-generating archives are steadily emerging. On these premises, it proposes a vision of archives as trustworthy producers of knowledge in the age of AI, one in which the knowledge archives generate is not restrained but made verifiable by disclosing its grounds and limits.

From theory to practice: realising the archival paradigm shift

In this paper, the knowledge that archives generate refers to meaningful statements constructed on the basis of records. The vision proposed here, that archives combined with AI can and should become major producers of trustworthy knowledge, presupposes that traditional archives did not systematically generate knowledge in response to users’ questions. Since the late twentieth century, however, archival science has regarded the archive as a site where the meanings of records are continually reconstructed; at the level of archival theory, a paradigm shift has already taken place. To maintain the premise,

therefore, this section first shows that, despite this theoretical shift, archives have continued to operate by providing users with pre-written descriptions and fixed records. It then argues that the knowledge-generating archive realises in practice the shift that occurred in theory, and draws on Assmann's theory of memory to explain its dual character of preservation and generation.

Modern archival science originally understood the archive as a repository of fixed evidence. In the classical view represented by Hilary Jenkinson, records are the natural by-products of the conduct of business, and their evidential value is maintained only when they are preserved intact in the state in which they were created. The archivist, as a neutral custodian who does not intervene in the records, is charged with safeguarding them as they were at the moment of their creation. From this perspective, the records of an archive are already fixed at the point of their creation and transfer, and the archive is no more than a vessel that transmits these fixed records unchanged (Jenkinson, 1937; Cook, 1997).

This view began to be revised in the mid-twentieth century, when T. R. Schellenberg established appraisal and selection by the archivist as a core task of archival science (Schellenberg, 1956), and it was fundamentally challenged from the late twentieth century onwards. Looking back over 150 years of archival thinking, Terry Cook traced four shifting paradigms, from evidence to memory to identity and community, in which archivists moved from passive keepers of records to active shapers of the archival heritage (Cook, 2013). One of the key theoretical contributions to this shift was Frank Upward's records continuum model, developed in Australia in the 1990s, which challenged the life-cycle model's assumption that the life of a record ends with its creation and transfer to the archive (Upward, 1996). Eric Ketelaar likewise observed that new meanings accrue to records each time they are used, arguing that records contain tacit narratives that need to be deconstructed (Ketelaar, 2001).

Indeed, as Duff and Harris (2002) showed, archival description is itself an interpretive act that narrates records and constructs their meaning, and records can be said to be always in the process of becoming. Yet description demands considerable time and labour, so that many archives struggled even to complete initial description (Greene & Meissner, 2005), and rewritten descriptions, too, were offered to all users in the same form. Models of the participatory archive, in which users annotate or take part in description, were also proposed (Huvila, 2008), but they were often pursued as experiments at a distance from the archive's own systems and struggled to sustain user involvement (Benoit & Eveleigh, 2019, p. 3). Reference services, in which archivists answer users' individual questions, are inherently limited in scale because they are performed directly by people, and requests calling for complex narratives rather than short answers require so much time and expertise that they are difficult to offer to all users. In short, although the theoretical shift proposed by Cook and others recognised not only archivists but also users and communities as active agents of meaning-making, responding to individual questions and bringing fixed records to life never became a systematic function of the archive; it remained largely the task of users themselves.

In the age of AI, the work that users once largely performed outside the archive, namely constructing the meaning of records in response to questions and thereby generating knowledge, enters the archive itself.

At the same time, the archive remains a place that preserves fixed records and their descriptions. The dual character of the knowledge-generating archive, which thus combines preservation and generation, can be explained through Aleida Assmann's theory of cultural memory. Assmann distinguishes functional memory (Funktionsgedächtnis), which a group activates in the present to create identity and meaning, from storage memory (Speichergedächtnis), the totality of latent memories kept but not currently in use, and associates the former with the canon and the latter with the archive (Assmann, 2011, p. 149). Although the appraisal and description of records already entail the construction of meaning, the traditional archive can still be placed in the position of storage memory, because what it produces remains largely latent until activated by users who bring their questions to it. Apart from the limited responses offered through reference services, the work of reading records and giving them meaning as present knowledge has for the most part been carried out beyond the archive, by historians and the many other users who interpret records for themselves. The knowledge-generating archive brings this activation into the archive itself: when a question is posed, part of the storage memory is called up and activated in the form of a response. Admittedly, a transient response to an individual user is not the collective functional memory Assmann describes. Yet in activating storage memory according to present questions, it is structurally similar to the process by which storage memory becomes functional memory.

This structural similarity brings another of Assmann's insights to the fore. Assmann held that storage memory is not only a resource for future functional memory but also a basis for correcting functional memory when it becomes rigid or distorted (Assmann, 2011, p. 132). In the knowledge-generating archive, this corrective relationship is built into the system. Generated knowledge is designed to be traced back to the storage memory from which it was called up, that is, to the fixed original records and their metadata. Whereas LLMs that answer from parametric knowledge alone rest on an informal archive stripped of provenance and so cannot make such corrections (Desai et al., 2024), the knowledge-generating archive, which combines storage memory and functional memory within a single system, allows generated knowledge to be verified. The knowledge-generating archive thus continues the tradition of the archive in preserving fixed records, while differing clearly from it in that meaning is constructed in response to questions within the archive itself, together with its grounds. The next section illustrates this difference through a comparison with an LLM trained on the same records.

Absorbed or preserved: comparing a trained language model and a knowledge-generating archive

There are two broad ways of combining archival materials with AI. One is to take the materials out of the archive and use them to train a language model, absorbing them into the model itself; the other is to leave the materials outside the model and retrieve them as grounds whenever a question is asked. In the terms of Lewis et al. (2020), the former turns the archive into parametric memory and the latter into non-parametric memory. The knowledge-generating archive belongs to the latter. This case study examines how the knowledge generated differs when the same materials are combined with AI in these two ways.

Experimental design

Materials. The study used the 517 letters retrieved by searching for Sim Neunggyu (□□□) in [a Korean studies digital archive, anonymised for review]. Of these, 470 are recorded as received by Sim Neunggyu; among those 470, the sender is recorded as unknown for 38 and as inferred for 5. Each letter carries an image of the original, a transcribed text, a summary (haeje), a scholarly note that explains its content and significance, and metadata such as sender, recipient, date and place, structured in XML. The materials are, however, incomplete. The letters have not been translated, the summaries vary considerably in detail, the sender or recipient field is sometimes empty, and the accuracy of the metadata has not been fully verified. The materials were used without cleaning, as served at the time of the experiment (October 2026), in order to reflect the conditions under which most archives would deploy AI: on top of incompletely described records. The same snapshot was supplied to both conditions.

Models. Both conditions were built on the same base models, Gemma 4 31B-it and Qwen3.8-27B, so that differences in results would arise not from the models but from the way the materials were combined with them. These are open-weight models that can be trained and run within the institution, which made it possible to build both conditions on the same model and allows the experiment to be reproduced.

Condition A

the trained language model. The base model was further trained on the metadata, summaries and transcribed classical Chinese texts of the 517 letters using LoRA adapters, a method that trains small added components rather than the whole model and so avoids degrading its general language ability. Several checkpoints from the course of training were tested (after 480, 640, 720 and 800 training steps for Gemma, and after 560 steps for Qwen), together with the base model before training. The model answered from memory alone, without access to the records, and was asked to tag each sentence with the type of grounds on which it rested (holdings, background explanation or inference).

Condition B: the knowledge-generating archive. The base model, without further training, answered by using the database as a set of tools: metadata lookup, a relational graph built from the metadata and the summaries, vector search, and reading the transcribed classical Chinese text of individual records. Responses were produced sentence by sentence, each tagged with its type of grounds and the identifiers of the records it cited.

Requests and runs. Both conditions received two requests. The first asked for the people who sent letters to Sim Neunggyu, together with the number of letters, in order from those who sent the most. The second asked for a narrative of about 5,000 characters reconstructing his life on the basis of the letters he received. This length was deliberately set beyond what the metadata and summaries of the collection could support. The aim was not to obtain a narrative of that length but to observe how each condition behaves when asked for more than its records can ground: whether it fills the gap with invented content or stops where the records end. A revised version of the second request, which asked the model to connect the events in the letters as chronological scenes rather than as lists of names and counts, was

also run for Condition B. Each configuration was run six times: once with greedy decoding (temperature 0), which always selects the most probable output, and five times with sampling (temperature 0.7, top-p 0.95), which introduces variation between runs. This made it possible to observe not only each condition's most probable answer but also how stable its answers were. Web search was disabled throughout.

Evaluation. Since the accuracy of the metadata has not been fully verified, what the experiment assesses is not whether responses agree with historical fact but whether they are faithful to the records the archive holds. The reference answer for the first request was obtained by counting, by sender, the letters received by Sim Neunggyu in the experimental database. Responses were assessed for agreement in ranking and counts, the presence of people who do not appear in the records as senders, the treatment of letters with unknown senders, and whether the cited records corresponded to the statements they were attached to.

Results

The first request

ranking of senders. In the reference answer, the five most frequent senders were Ryu Yeongha (□□□, 29 letters), Kim Nogyeom (□□□, 27), Kim Illyeon (□□□, 21), Im Jeongsang (□□□, 17) and Yi Johyeong (□□□, 15).

The knowledge-generating archive placed Ryu Yeongha first with 27 letters in all twelve runs across both models, despite the variation introduced by sampling, and the order it gave never departed from the one recorded in the archive. All six Gemma runs and three of the six Qwen runs gave the full top five; the other three Qwen runs stopped after listing two to four senders. The counts it gave (27, 24, 20, 15 and 12) were lower than those in the reference answer for most senders, because the database records some individuals as several person entries owing to differences in recorded dates of birth and death, and the archive's counting tool counts by entry. Kim Nogyeom, for example, appears as three entries with 24, 2 and 1 letters, and the archive reported the largest. Hong Jaecheol (□□□) likewise appeared twice in several responses, with 8 and 6 letters; one Qwen response noted that two entries under this name together accounted for 14 letters. The discrepancy is thus not a hallucination but a faithful reflection of an inconsistency in the archive's own description, and one that a user can trace and correct.

The trained language model did not produce the correct ranking in any run, and its answers changed from run to run. Before training, the Gemma base model named well-known figures of the Joseon period, such as Yi Hangbok, Yi I and Jeong Yakyong, as the main senders and attributed letter counts to them; none of them appears in the collection. After training, the names it produced came from the collection, but their roles were wrong. In 16 of the 24 runs across the four Gemma checkpoints, the model placed Sim Wontak (□□□) first, describing him as Sim Neunggyu's son and attributing 7 to 10 letters to him. In the database, Sim Wontak sent no letters to Sim Neunggyu; he appears as the recipient of eight. Ryu Yeongha, the actual leading sender, was placed first in only three runs, with 6 to 10 letters rather than 29. The Qwen base model, by contrast, declined to answer in all six runs, stating that it had no information

on the senders. After training, the Qwen checkpoint declined in three runs and in the other three named one or two people with one or two letters each (Table 1).

Table 1

Responses to the first request. The reference answer gives Ryu Yeongha as the leading sender with 29 letters.

Configuration	Runs	Leading sender given	Correct top five
A · Gemma base model	6	Well-known figures absent from the collection (6 runs)	0
A · Gemma trained (4 checkpoints)	24	Sim Wontak, who sent no letters (16 runs); Ryu Yeongha with 6–10 letters (3 runs); others (5 runs)	0
A · Qwen base model	6	Declined to answer (6 runs)	0
A · Qwen trained	6	Declined (3 runs); one or two people with one or two letters (3 runs)	0
B · Gemma	6	Ryu Yeongha, 27 letters (6 runs)	6
B · Qwen	6	Ryu Yeongha, 27 letters (6 runs)	3 (the other 3 stopped after 2–4 senders)

Two limitations of the knowledge-generating archive also emerged in this request. The first was an omission. The counting tool returned the 38 letters with unknown senders in every run, so the information was available to the model, yet none of the twelve responses mentioned them. The responses were not wrong, but they did not convey to the user the uncertainty the archive knew about. The second concerned citation. The records attached to each sentence did not correspond to its content. In the Gemma responses, the same 80 records were attached to every sentence, and of the 80 records attached to the sentence naming Ryu Yeongha as the leading sender, none was a letter sent by him. Some Qwen responses attached as many records as the count they reported, but of the 27 records attached to the sentence about Ryu Yeongha, again none was sent by him. The counts themselves were faithful to the records, because they came from querying the database directly, but the citations offered as their grounds did not support them.

The second request

a narrative reconstructing a life. None of the completed responses reached the requested length of about 5,000 characters. The responses of the trained model that exceeded this length were all cut off at the output limit, often after falling into repetition. The completed responses of the trained model ranged from about 1,000 to 4,700 characters, and those of the knowledge-generating archive from about 700 to 2,900 (Table 2).

The narratives of the trained model were fluent but drew largely on the typical life of a Joseon literati scholar rather than on the letters. One response, for example, described Sim Neunggyu as passing the

civil service examination in the 1830s, serving as a local magistrate and then in high central office in the 1840s and 1850s, and retiring to his hometown, none of which is supported by the records. Biographical details varied from run to run, including his dates of birth and of passing the examinations, his clan seat and the titles of his writings, and the model presented some of these under the tag of holdings. The trained Gemma checkpoints cited record identifiers in the correct format, and most of them existed in the collection (for example, 437 of the 560 identifiers cited by the checkpoint after 480 steps), but whether the cited records supported the sentences was not examined. The Gemma base model, before training, cited invented identifiers.

The narratives of the knowledge-generating archive were shorter, and those produced with the original request tended towards lists of senders and dates. With the revised request, they took the form of dated scenes linked to specific letters. One Qwen response, for example, traced Sim Neunggyu from his birth in 1790 and his adoption into another branch of the family, through letters received from 1813 onwards, to his passing the licentiate examination in 1859 at the age of seventy, and noted explicitly that there is no record of his holding office. Almost every sentence was tagged as grounded in the holdings. An informal check by the author of a sample of 30 sentences found that the records cited in these narratives broadly corresponded to the content of the sentences, in contrast to the first request; a systematic sentence-by-sentence verification, however, was not carried out.

Table 2
Responses to the second request (requested length: about 5,000 characters).

Configuration	Completed length	Character of the narrative
A·trained models	about 1,000–4,700; longer outputs cut off at the output limit	Fluent; careers, offices and biographical details not supported by the records; details vary across runs
B·original request	about 700–1,100	Short; tends towards lists of senders and dates
B·revised request	about 1,000–2,900	Dated scenes linked to specific letters; absence of records stated where relevant

Discussion

Who sent how many letters is an arbitrary fact in the sense of Kalai et al. (2025): it cannot be inferred from general patterns and is confirmed only by the records. The case study shows what happens to such facts when they are absorbed into a model’s parameters. Training did not teach the model the correct ranking, at any of the checkpoints tested; it changed the kind of error the model made. Before training, the Gemma model filled the gap with famous figures from outside the collection. After training, it filled the gap with names from inside the collection, but in the wrong roles, turning a recipient into the leading sender. The error became more plausible, and therefore harder to detect, precisely because the model had learned the collection. That this held regardless of how far training had progressed suggests that the problem lies not in insufficient or excessive training but in parametric memory itself. In the

knowledge-generating archive, by contrast, the same information remained in a structure that could be counted directly, and the archive never departed from the recorded order in any run, however much variation sampling introduced.

The experiment also shows that the knowledge-generating archive is not free from error, and the nature of its errors is instructive. Its counts diverged from the reference answer where the archive's own description split one person into several entries; this is an inconsistency in the records, but one that the response exposed rather than concealed. The real failures lay elsewhere. It retrieved the letters with unknown senders in every run but never disclosed them, and its correct counts were attached to records that did not support them. The latter is the problem Magesh et al. (2025) identified in commercial legal research tools: presenting sources is not the same as grounding a statement in them. The knowledge-generating archive held the means of verification, in the form of records and their provenance, but its responses did not yet make proper use of them. These failures show that the principle of verification cannot be left to the language model. It must be built into the archive as procedures that ensure the uncertainty the archive knows about is disclosed, and that each cited record actually supports the statement it is attached to. In the narratives, by contrast, the cited records broadly supported the sentences, which suggests that the problem lies less in the archive's capacity to ground statements than in how aggregated answers are linked to individual records.

The narratives point in the same direction. Since the requested length exceeded what the records could support, the way each condition responded to this excess is itself revealing. The trained model produced fluent life stories in which what the letters record and what was invented could not be told apart, and its longest outputs ran on until they broke off in repetition. The knowledge-generating archive produced shorter accounts that stayed close to dated letters and stated where the records were silent, as in the observation that no record shows Sim Neunggyu holding office. Its narratives also depended on how the request was framed, which suggests that how the archive constructs knowledge in response to a question is itself a matter of design.

This case study rests on the letters received by a single individual and on two requests, and the correspondence between the narratives and the records they cite was checked only informally. Moreover, both conditions had access to the transcribed classical Chinese texts of the letters, which have not been translated, but how the language models interpreted these texts, and whether their interpretations were accurate, was not evaluated separately in this study. Comparisons with providing the same materials directly in a model's context window, or with other types of records, are left for future research. Even so, the case study shows that archives can contribute to the knowledge ecosystem of the age of AI as trustworthy producers of knowledge, provided that verification is built into them. The section that follows sets out the vision of the knowledge-generating archive through three principles: the wellspring, verification and generation.

A vision for archives in the age of AI

Archival science has long treated the reliability and authenticity of records as its core values (Duranti, 1995). These values can be understood as preconditions that the archive guarantees so that users can rely on them when creating knowledge outside the archive. The wellspring and verification correspond to these traditional values of archival science. Yet the vision of the knowledge-generating archive does not merely restate values that archival science has long emphasized. The wellspring captures the archive's role in supplying sources to a knowledge ecosystem in which LLMs have become powerful actors, and verification captures its role in presenting the grounds of the knowledge it generates. Generation, proposed alongside these two, is a new principle that positions the archive as a significant actor in the knowledge ecosystem. This section examines the wellspring, verification and generation in turn, and then sets out the relationship among the three principles.

The wellspring

The significance of the wellspring in the age of AI becomes clear in a study published in Nature in 2024 by Ilia Shumailov and his colleagues (Shumailov et al., 2024). They asked what happens when text produced by large language models comes to make up a substantial part of the internet. If future AI models are also trained on data collected from the web, they will inevitably end up learning from data produced by earlier models. Their experiments showed that models trained indiscriminately on AI-generated data forgot, over successive generations, the distribution of the original data, that is, the pattern of how often different kinds of content appear in the original material. The authors called this 'model collapse' and showed that, when such indiscriminate retraining continues, it leads to a degeneration that is difficult to reverse, a phenomenon that appears not only in language models but in generative models in general.

The pattern of model collapse is telling. Collapse begins with the loss of the tails of the distribution, that is, of rarely occurring events and expressions, and ends with all outputs converging on a narrow range that bears little resemblance to the original. The authors stress that the ability to model rare events is essential for fairness, since such events are often associated with marginalised groups. The letters of a Joseon-period literati scholar or the old documents of a local lineage are precisely such 'tails' within the vast body of web data. The more a model is trained on data it has itself produced, the sooner this kind of rare and specific knowledge is erased. The authors conclude that sustaining the training of models over the long term requires preserving access to original data sources and ensuring that data not generated by language models remain available.

Subsequent research has shown that collapse can be avoided when new data are accumulated alongside existing data rather than replacing them (Gerstgrasser et al., 2024). On this view, model collapse is not an inevitable fate but a problem that can be overcome, provided that the original human-produced data are preserved rather than replaced. As AI-generated text increasingly fills the knowledge ecosystem, then, the value of records that can be guaranteed to have been produced without the intervention of AI only grows. To this end, archives need to distinguish and record which of their holdings were produced by people and which were created with the help of AI. In an age in which LLMs are

emerging as major producers of knowledge, the wellspring as a principle of the knowledge-generating archive goes beyond the merely passive sense of keeping old records. Archives holding records produced by people will serve as the wellspring, the source that continually supplies life to the knowledge ecosystem of the age of AI.

Verification

Hallucination in LLMs is not an accidental defect but a statistical consequence of the way language models are trained. Kalai et al. (2025) showed that, for facts that exist individually in the training data without any consistent pattern, which they call ‘arbitrary facts’, such as a particular person’s birthday, pretrained base models can hardly avoid errors. On their analysis, the proportion of facts of a given kind that appear only once in the training data sets a lower bound on the hallucination rate: if 20 percent of birthday facts appear only once in the training data, a base model can be expected to hallucinate on at least 20 percent of questions about birthdays. They attribute the persistence of such hallucinations to training and evaluation procedures that reward guessing over acknowledging uncertainty. The senders of the letters used in the case study above, and the number of letters each sent, are likewise arbitrary facts. They cannot be inferred from general patterns and can be confirmed only through the records themselves.

As noted in the Introduction, retrieval-augmented generation (RAG) emerged as a technical response to hallucination (Lewis et al., 2020). RAG, however, does not eliminate hallucination entirely. Magesh et al. (2025) evaluated commercial RAG-based tools that claimed to offer hallucination-free legal research in a pre-registered study, and found that although these tools hallucinated less than general-purpose chatbots, they still did so in 17–33 percent of queries (Magesh et al., 2025, p. 216). To take the requests in the case study as an example, a request that can be answered from metadata, such as ‘Show the people who sent letters to Sim Neunggyu, together with the number of letters, in order from those who sent the most’, differs from a request that requires the generation of a narrative, such as ‘Write a narrative reconstructing Sim Neunggyu’s life on the basis of the letters he received’. In the latter case, hallucination can occur even in the knowledge-generating archive, because the language model fills the gaps the records leave with what it has learned.

Overcoming this problem requires not only indicating, for each sentence of a narrative, the type of grounds on which it rests, such as holdings or inference, but also making it possible to check whether the passage of the record cited as grounds actually supports that sentence. Many current LLMs do, of course, present their sources, but what they search is material whose reliability is not guaranteed. The knowledge-generating archive, by contrast, holds provenance metadata organised by experts, together with the original records. Unlike LLMs, in other words, it carries within itself firm grounds for verification. As the case study showed, however, holding the grounds for verification is not the same as using them: the archive’s correct counts were attached to records that did not support them, and the uncertainty it had retrieved went undisclosed. Verification must therefore be built into the archive as a procedure, not left to the language model. Grounds for verification, so used, also answer Finn and Khosrowi’s concern

that the mediated nature of sources may be obscured and source criticism bypassed (Finn & Khosrowi, 2026). The remedy for their concern, that is, lies not in restraining the archive's capacity to generate knowledge but in ensuring that generated knowledge discloses its own grounds and limits. When grounds and limits are disclosed, users can check the grounds and accept generated knowledge not as a final interpretation but as one among several possible interpretations. In such circumstances, source criticism and interpretation still rest with users. The principle of verification is thus also a way of fulfilling the responsibility of the archive that Finn and Khosrowi's concern presupposes.

Generation

Whereas the wellspring and verification are values long upheld by traditional archives, transformed in response to the demands of the age of AI, generation is a newly required principle. As discussed earlier, traditional archives have also provided knowledge through catalogues and descriptions, but this was for the most part ready-made knowledge, written once at the time of acquisition and offered in the same form to all users. Records have occasionally been redescribed with the participation of users, but such arrangements can hardly be said to have been common. Users could not obtain from the archive knowledge processed into the form they wanted, and the work of reading and connecting records to construct answers to their questions was in practice left to users themselves. Generation in the knowledge-generating archive refers to bringing this process into the archive itself, constructing grounded knowledge in response to users' requests. Knowledge generated in this way has the following three characteristics.

The first is context-dependence. Knowledge in the knowledge-generating archive is constructed anew in response to each user's question. The same letter is called up in different contexts depending on whether the question concerns the life of its sender or the circumstances of a particular event, and the knowledge graph links the letter to the relevant persons and events to construct the context the question requires. The second is variability. When new source materials, new research findings or new interpretations are added, the knowledge graph and index are updated, and the response to the same question changes accordingly. Knowledge is not fixed at the point of transfer but continues to grow throughout the life of the archive. The third is verifiability. The first two characteristics can be risky in themselves: knowledge that is constructed anew for each question and constantly changing becomes, the moment it loses its grounds, no different from the fluent guesswork of general-purpose language models, that is, from hallucination. Statements generated in the knowledge-generating archive must therefore always be traceable back to the original records and metadata on which they are based.

The relationship among the three principles

In Assmann's terms, generation corresponds to activating storage memory as functional memory in response to questions, and verification to returning activated functional memory to the storage memory from which it was called up, so that it can be corrected. Verification and generation can thus be regarded as internal principles of the archive itself, each dependent on the other. Unlike these two, the wellspring can be regarded as an external principle through which the archive keeps the knowledge ecosystem

healthy. Yet this external principle is not unrelated to the internal ones. Knowledge is generated by weaving together records and their descriptions; generated knowledge is verified by returning to the sources on which it rests; and grounded knowledge, flowing outward while still pointing back to the preserved originals, helps sustain the robustness of the knowledge ecosystem. Generation, verification and the wellspring thus form an organic system, each sustaining the others.

Conclusion

Writing and paper, printing and digital technologies profoundly changed the ways in which knowledge was recorded and disseminated, and these changes in turn deeply affected the ways in which knowledge was produced. Yet these technologies only assisted the production of knowledge; its producers were always human beings. AI differs from earlier technological changes in that it has emerged as an actor that generates knowledge itself. The emergence of this powerful new producer of knowledge portends significant change for archives as well. Paradoxically, RAG, proposed to overcome hallucination as a fundamental limitation of AI, has become the driving force for realising in practice the paradigm shift of the archive in the age of AI. In this situation, this paper has asked what vision archives should hold, and has proposed three principles: generation and verification, interdependent internal principles of the archive, and the wellspring, an external principle that sustains the knowledge ecosystem.

The wellspring and verification are also traditional values of the archive, but they take on new meaning in the age of AI. The wellspring requires that, in the face of the risk of model collapse, the archive breathe life into the knowledge ecosystem by supplying sources produced by people and of assured provenance. Verification requires that it be possible to check whether generated knowledge is anchored in its grounds. Generation captures the archive's role in bringing into itself the construction of knowledge that once took place outside it, and in generating knowledge whose grounds and limits are made explicit. The three principles sustain one another. Without the wellspring, the knowledge ecosystem to which the archive belongs is put at risk of collapse; without verification, generation cannot be distinguished from hallucination; and without generation, the role of the wellspring remains limited. The contribution of this paper lies in repositioning the traditional values of archival science to fit the new functions of the archive in the age of AI, and in combining them with the principle of generation into a single system.

This paper also has limitations. The case study is confined to 517 letters received by a single individual, and further research is needed to generalise its findings to other types of records or to larger collections. The questions, too, need to be diversified by type; it would be worth comparing, for example, responses to complex questions such as a request to evaluate two conflicting interpretations. Not only the questions but also the systems compared need to be diversified. In particular, this paper compared only a trained LLM with a knowledge-generating archive, and did not include an approach in which the same materials are provided directly to the context window of a language model. This limitation is likely to invite the objection that, as context windows grow larger, there may be no need to build a knowledge-generating archive at all.

This objection can be answered in three ways. First, while one individual's letters could fit within a context window, supplying an institution's entire holdings with every question is impractical in scale and cost. Second, an expanded context window can replace the technique of retrieval, but not the archive that creates and safeguards trustworthy sources; the more a model can read at once, the more important it becomes to guarantee by whom and in what context that material was produced, and what it does not tell us. Third, a knowledge-generating archive structures its materials through a knowledge graph, making explicit the relationships among persons, places, events and records that remain implicit in raw text. Whether a question asks for counts or for a narrative of a life, the archive can follow these relationships across records and assemble a response whose links can be traced back to its sources. Nonetheless, these and other questions arising from the paper's limitations require empirical examination, which would also give concrete form to the vision proposed here; we leave this to future research.

Archives in the age of AI must not stop at preserving fixed records; they must anchor the knowledge they generate so that it does not drift, while creating new knowledge together with their users. In this, they will come to resemble laboratories. Just as hypotheses are formulated and tested against evidence in a laboratory, in the knowledge-generating archive the user's question becomes a hypothesis, the generated response a provisional answer to it, and the preserved original records the evidence against which that answer is tested. A laboratory that keeps the wellspring, verifies knowledge and generates it: this is the vision of archives in the age of AI.

Declarations

Author Contribution

W.J. conceived and designed the study, interpreted the results and wrote the main manuscript text. J.A. and G.L. built the knowledge-generating archive and the trained language models and conducted the experiments. I.O., C.K. and J.L. prepared the archival materials and contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

Data Availability

The letters analysed in this study are publicly accessible through [a Korean studies digital archive, anonymised for review]. The experimental outputs generated during this study, including the configuration of each run, the full responses of both conditions and the run provenance records, are provided as Supplementary Information for peer review and will be deposited in a public repository upon acceptance. The snapshot of the metadata, summaries and transcribed texts used in the experiment, together with the prompts and the reference answers, is available from the corresponding author on reasonable request.

References

1. Assmann A (2011) Cultural memory and Western civilization: functions, media, archives. Cambridge University Press, Cambridge
2. Barbudo Carrasco, Engvall T, Dobslaw F (2026) A conceptual crosswalk between trustworthy records and trustworthy AI. Philos Technol 39:100. https://doi.org/10.1007/s13347-026-01112-6
3. Benoit E, Eveleigh A (eds) (2019) Participatory archives: theory and practice. Facet, London
4. Burton JW, Lopez-Lopez E, Hechtlinger S et al (2024) How large language models can reshape collective intelligence. Nat Hum Behav 8:1643–1655. https://doi.org/10.1038/s41562-024-01959-9
5. Colavizza G, Blanke T, Jeurgens C, Noordegraaf J (2021) Archives and AI: an overview of current debates and future perspectives. J Comput Cult Herit 15(1):4:1–415. https://doi.org/10.1145/3479010
6. Colavizza G, Jaillant L (2026) AI ready: how to prepare archives for artificial intelligence, improve discoverability, and enhance access. Arch Rec. https://doi.org/10.1080/23257962.2026.2675937
7. Cook T (1997) What is past is prologue: a history of archival ideas since 1898, and the future paradigm shift. Archivaria 43:17–63
8. Cook T (2013) Evidence, memory, identity, and community: four shifting archival paradigms. Arch Sci 13(2–3):95–120
9. Davies P (2024) An AI tool is making the European Parliament's history easier to understand in all EU languages. Euronews, 22 October 2024. https://www.euronews.com/next/2024/10/22/an-ai-tool-is-making-the-european-p arliaments-history-easier-to-understand-in-all-eu-langu
10. Desai MA, Pasquetto IV, Jacobs AZ, Card D (2024) An archival perspective on pretraining data. Patterns 5(4):100966. https://doi.org/10.1016/j.patter.2024.10096
11. Duff WM, Harris V (2002) Stories and names: archival description as narrating records and constructing meanings. Arch Sci 2(3–4):263–285
12. Duranti L (1995) Reliability and authenticity: the concepts and their implications. Archivaria 39:5–10
13. Duranti L (2005) The long-term preservation of accurate and authentic digital data: the InterPARES project. Data Sci J 4:106–118. https://doi.org/10.2481/dsj.4.106
14. Edge D, Trinh H, Cheng N et al (2024) From local to global: a graph RAG approach to query-focused summarization. https://doi.org/10.48550/arXiv.2404.16130. arXiv
15. Finn F, Khosrowi D (2026) AI assistants in the archive and the lure of instant history. Camb Forum AI Cult Soc 2:e6. https://doi.org/10.1017/cfc.2025.10012
16. Frontoni E, Paolanti M, Lauriault TP, Stiber M, Duranti L, Abdul-Mageed M (2022) Trusted data forever: is AI the answer? arXiv. https://arxiv.org/abs/2203.03712
17. Gerstgrasser M, Schaeffer R, Dey A, Rafailov R, Sleight H, Hughes J, Korbak T, Agrawal R, Pai D, Gromov A, Roberts DA, Yang D, Donoho DL, Koyejo S (2024) Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. https://doi.org/10.48550/arXiv.2404.01413. arXiv

18. Greene MA, Meissner D (2005) More product, less process: revamping traditional archival processing. Am Arch 68(2):208–263
19. Huvila I (2008) Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management. Arch Sci 8(1):15–36
20. Jaillant L (ed) (2022) Archives, access and artificial intelligence: working with born-digital and digitized archival collections. transcript, Bielefeld. https://doi.org/10.14361/9783839455845
21. Jaillant L, Rees A (2022) Applying AI to digital archives: trust, collaboration and shared professional ethics. Digit Scholarsh Humanit. https://doi.org/10.1093/llc/fqac073
22. Jenkinson H (1937) A manual of archive administration, new and revised edn. Percy Lund, Humphries & Co, London
23. Jo ES, Gebru T (2020) Lessons from archives: strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, p 306–316. https://doi.org/10.1145/3351095.3372829
24. Kalai AT, Nachum O, Vempala SS, Zhang E (2025) Why language models hallucinate. arXiv. https://doi.org/10.48550/arXiv.2509.04664
25. Ketelaar E (2001) Tacit narratives: the meanings of archives. Arch Sci 1(2):131–141
26. Koebler J (2024) 'AI-mazing tech-venture': National Archives pushes Google Gemini AI on employees. 404 Media, October 2024. https://www.404media.co/ai-mazing-tech-venture-national-archives-pushes-google -gemini-ai-on-employees/
27. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. https://doi.org/10.48550/arXiv.2005.11401. arXiv
28. Luxembourg Government (2023) Eluxemburgensia.lu now has a new chatbot: artificial intelligence at the service of Luxembourg's printed heritage. Press release, 4 October 2023. https://gouvernement.lu/en/actualites/toutes_actualites/communiques/2023/10-octobre/04-tanson-eluxemburgensia.html
29. Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE (2025) Hallucination-free? Assessing the reliability of leading AI legal research tools. J Empir Leg Stud 22(2):216–242. https://doi.org/10.1111/jels.12413
30. Schellenberg TR (1956) Modern archives: principles and techniques. University of Chicago Press, Chicago
31. Shinde G, Kirstein T, Ghosh S, Franks PC (2024) AI in archival science: a systematic review. arXiv. https://doi.org/10.48550/arXiv.2410.09086
32. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y (2024) AI models collapse when trained on recursively generated data. Nature 631:755–759. https://doi.org/10.1038/s41586-024-07566-y

33. Upward F (1996) Structuring the records continuum part one: postcustodial principles and properties. Arch Manuscr 24(2):268–285

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData1READMEEnglish.md
- SupplementaryData1Experimentaloutputs.zip