

# A Study on Archaeological Informatization Using Large Language Models (LLMs)

## - Proof of Concept for an Automated Metadata Extraction Pipeline from Archaeological Excavation Reports -

**KIM Hongyeon**

*Undergraduate student, Dept. of Computer Science, Korea National Open University;  
B.A. in Archaeology, Korea National University of Heritage*

### Abstract

The field of archaeology handles vast quantities of data, much of which exists in unstructured or semi-structured textual forms, which poses persistent challenges to its systematic use and dissemination. Previous efforts at informatization have often failed to fundamentally improve the research environment or data accessibility due to the unique characteristics of archaeological data and its ever-increasing volume. This paper first systematically reviews various past discussions on archaeological informatization against this backdrop and proposes a new strategy to approach this problem using large language models (LLMs), which have recently emerged as an innovative technology. Specifically, it explores the potential of using LLMs for automatic metadata extraction and information structuring from documents such as excavation reports, which often interweave repetitive narrative structures with standardized informational elements. Through this, it investigates the possibilities for improving research efficiency as well as the multifaceted use of archaeological knowledge in academic research, exhibitions, education, and cultural content development.

In particular, the proof of concept (PoC) results of an LLM-based informatization application developed by the author serve as crucial evidence, going beyond a simple technology introduction, to propose a practical direction for redesigning the core structure of archaeological informatization in a more realistic and flexible manner. Furthermore, it suggests the potential for expanding knowledge services using LLMs post-informatization and highlights the potential for data-driven analysis to become a routine tool in archaeological research. Ultimately, this study presents new possibilities for advancing archaeology by applying LLMs as a core technology to revolutionize the overall methods of knowledge production, access, and interpretation. For successful archaeological informatization, it is important to secure source digital data and establish an institutional foundation for its systematic management and disclosure. The theoretical and technological foundations proposed herein may also be a vital starting point for future discussions on the digital transformation of cultural heritage in general.

**Keywords:** Artificial Intelligence (AI), Large Language Models (LLM), Excavation Reports, Digital Archaeology, Metadata

**Received** May 25, 2025 | **Revised** July 15, 2025 | **Accepted** August 6, 2025 | **Translated** April 14, 2026

---

*Currently affiliated with heripo lab*

ORCID: <https://orcid.org/0009-0000-6955-3023> | Correspondence: [kimhongyeon89@gmail.com](mailto:kimhongyeon89@gmail.com)

This is a translated version of the following article: KIM Hongyeon, 2025, "대형 언어 모델(LLM)을 활용한 고고학 정보화 연구 – 발굴조사보고서의 메타데이터 자동 추출 파이프라인 개념 검증 –," *Korean Journal of Heritage: History & Science*, 58(3), pp. 34–61. DOI: [10.22755/kjchs.2025.58.3.34](https://doi.org/10.22755/kjchs.2025.58.3.34)

This English edition was translated and edited by the author with the permission of the National Research Institute of Cultural Heritage, Korea. The translation includes corrections to factual errors identified in the original Korean edition and supplementary translation notes to aid international readers' understanding. These modifications do not affect the core arguments, results, or interpretation of the original paper.

Copyright © 2025 National Research Institute of Cultural Heritage, Korea.

## I. Introduction

As an undergraduate studying archaeology, I often found myself questioning whether archaeology was truly about studying culture, or merely about classifying data embedded in reports and records. Rather than scholarly activities such as research, critical examination, or analysis, what I experienced firsthand was a reality dominated by what could only be described as manual drudgery. At times, I felt deep skepticism upon witnessing cases where simple repetitive work—devoid of any meaningful analytical process—was nonetheless recognized as scholarly achievement. These experiences drew my attention to the possibility that digital technologies could dramatically improve research efficiency and analytical precision by enhancing access to archaeological data and lowering the barriers to information use. I also became convinced that if archaeological literature such as excavation reports were informatized, their value could extend far beyond mere archival storage into diverse fields including academic research, museum exhibitions, education, and cultural content development.

These concerns were first crystallized in my undergraduate thesis, where I argued that “information acquired digitally should be expressed digitally,” emphasizing the need for structural reform and institutional groundwork to move beyond analog-centered recording systems (Hongyeon Kim 2013: 285). To explore this issue further at the policy level, I enrolled in a master’s program in public policy, during which I focused on analyzing the realities and limitations of data-driven cultural heritage administration. At the time, Republic of Korea (hereafter Korea) government-led initiatives such as open data policies and “Cultural Heritage 3.0” (문화유산 3.0) remained largely declarative, lacking a solid foundation for private-sector utilization or practical linkage to fieldwork (Hongyeon Kim 2015: 323–326).

I eventually reached the conclusion that no matter how well institutional frameworks were established, archaeological informatization could not function without genuine improvements in both the technical environment and the capacity of field practitioners. Even if institutional standards were put in place, the reality remained that a vast body of reports and literature had already accumulated without any structural organization. My focus naturally shifted from institutions and policy to technology itself. I subsequently transitioned into the IT industry, working as a software engineer and spending over a decade solving diverse problems through technology.

Even as I built up practical experience and technical expertise in the IT industry, archaeological informatization remained a central concern. I never let go of the goal of someday building practical tools for organizing and analyzing archaeological data, and this vision continued to take shape in my mind over the years.

More recently, with the widespread adoption of Large Language Models (LLMs),<sup>1</sup> exemplified by ChatGPT,<sup>2</sup> it became possible to automatically organize vast documents and unstructured information to a reasonable degree—tasks that had previously required entirely manual effort. New solutions began to emerge for the very problems I had long faced: disorganized reports, repetitive labor, and the difficulty of comparative analysis.

LLMs are more than simple response-generation tools; they are structuring tools that automate repetitive language tasks. They are particularly effective at processing semi-structured<sup>3</sup> texts containing repetitive data, such as excavation reports, offering a technology-driven alternative to the institutional and labor-intensive approaches that had previously failed to achieve archaeological informatization.

This paper synthesizes existing discussions on archaeological informatization and explores the potential for reconfiguring informatization strategies in light of recent digital advances, including LLM technology. Going beyond a simple technology introduction, this study develops a Proof of Concept (PoC)<sup>4</sup> application that uses LLMs to automatically extract metadata from excavation reports, thereby validating the feasibility of the concept. Furthermore, it examines additional possibilities built upon this foundation, such as natural language search, information summarization, and comparative analysis.

---

<sup>1</sup> [Translation note] The original footnote discussed the various Korean-language translations of “Large Language Model”—including “대형 언어 모델,” “대규모 언어 모델,” and “거대 언어 모델”—noting that no standardized Korean term had been established at the time of publication. As this footnote pertains exclusively to Korean-language terminology, it cannot be meaningfully rendered in English translation. (Original footnote 1)

<sup>2</sup> OpenAI’s ChatGPT was first released on November 30, 2022. It reached approximately one million users within five days of launch and approximately 100 million users by January 2023. “ChatGPT,” Wikipedia, accessed April 30, 2025, <https://en.wikipedia.org/wiki/ChatGPT>.

<sup>3</sup> Although archaeological literature such as excavation reports may appear to consist of unstructured narrative, it actually possesses the characteristics of semi-structured data, with recurring informational elements such as artifact dimensions, findspot locations, and page references to drawings and photographs.

<sup>4</sup> A Proof of Concept (PoC) refers to an early-stage experiment or prototype development aimed at demonstrating that a particular concept or theory can be feasibly implemented and demonstrates practical utility. It typically involves examining technical feasibility within a limited scope, serving as a foundational validation for subsequent full-scale development.

## II. Current State and Challenges of Archaeological Informatization: The Korean Context

### 1. The Concept and Necessity of Archaeological Informatization

The term “digital transformation” has recently gained currency in archaeology as well. The digital transformation of archaeology can be understood as an attempt to restructure the full spectrum of archaeological activities—fieldwork, research, and cultural utilization—on a digital technology foundation. This encompasses, for example, producing field records in standardized digital formats, routinely employing technologies such as LLMs, machine learning,<sup>5</sup> and deep learning<sup>6</sup> in the research process, and extending the resulting knowledge into public cultural content.

While such aspirations toward digital transformation are encouraging, they all too often remain at the level of abstract discourse, lacking concrete plans for implementation. A closer look reveals that much of what is presented under the banner of “digital transformation” is still confined to the initial stages of informatization,<sup>7</sup> falling short of any fundamental shift in approach.

Accordingly, rather than invoking the broad concept of “digital transformation,” this paper focuses on the more concrete and actionable notion of “archaeological informatization.” As defined here, archaeological informatization refers to the process of organizing and structuring archaeological literature and data using digital technologies, building an information ecosystem that interconnects and integrates these resources, and—crucially—premiering the entire process on automation. Archaeological informatization should not be limited to organizing the data that has accumulated to date; it must incorporate automated systems capable of stably and efficiently managing the archaeological data that will continue to be generated in the future.

Unlike fixed historical records such as the *Goryeosa* (高麗史, History of Goryeo) or the *Joseon Wangjo Sillok* (朝鮮王朝實錄, Annals of the Joseon Dynasty), archaeological literature constitutes dynamic data that is continuously generated through thousands of investigations each year (Table 1). To manage such repetitive and unstructured material, an automation-based informatization framework—rather than one-time manual organization—is essential.

---

<sup>5</sup> A subfield of artificial intelligence in which a computer learns patterns from data on its own—without being explicitly programmed with rules—and performs predictions or classifications. For example, it can be applied to automatically classifying site characteristics based on artifact morphology and findspot data, or to identifying topic-specific sentences in archaeological reports.

<sup>6</sup> A subfield of machine learning that uses artificial neural networks modeled on the structure of the human brain to automatically learn from large-scale data and handle complex tasks. For instance, it can be applied to automatically detecting traces of tumuli or archaeological features in high-resolution aerial photographs, or to recognizing handwriting in historical manuscripts and converting it into machine-readable text.

<sup>7</sup> According to Article 2, Paragraph 2 of the Framework Act on Intelligent Informatization (지능정보화 기본법), “informatization” refers to the production, distribution, or utilization of information to enable or enhance the efficiency of activities across all sectors of society. This can be understood as a precondition and foundation for digital transformation.

**Table 1. Number of Surface Surveys and Excavations and Associated Costs (Korea Heritage Service, e-Nara Index)**

		2020	2021	2022	2023	2024
<b>Surface Survey</b>	Cases	1,652	1,601	1,667	1,485	1,340
	Cost (100M KRW)	105	101	116	119	172
<b>Excavation</b>	Cases	2,642	2,731	2,648	2,254	2,310
	Cost (100M KRW)	2,735	2,914	3,027	3,058	2,629
<b>Total</b>	Cases	4,294	4,332	4,315	3,739	3,650
	Cost (100M KRW)	2,840	3,015	3,143	3,177	2,801

\* [Translation note] The 2024 excavation figures (number of cases and cost) published in the original Korean edition contained errors. They have been corrected in this translation based on the authoritative source (Korea Heritage Service, e-Nara Index).

Moreover, even with the advent of the digital transformation era, it would be difficult to immediately introduce and fully establish a standardized digital recording system at excavation sites. Even if such a system were adopted, the very nature of archaeology means that new types of artifacts and exceptional site conditions are constantly encountered during fieldwork. Archaeological data are inherently resistant to permanent standardization; they must always accommodate data that exceed existing standards.<sup>8</sup>

Archaeological informatization must therefore be designed as a system capable of automatically organizing and integrating not only the data that have already accumulated, but also the data that will continue to change and grow in the future. This constitutes the foundation for the digital transformation of archaeology and, more broadly, an essential undertaking that underpins the sustainability of archaeological research and its practical applications.

Furthermore, archaeological informatization is more than mere data maintenance; it serves as the starting point that makes the substantive achievements of digital archaeology possible. Digital archaeology encompasses a wide range of digital technologies—geophysical survey, 3D scanning, spatial information systems, visualization techniques, and digital public archaeology, among others (Choi 2022: 3)—yet none of these technologies can function properly without structured and systematically managed archaeological data. Ultimately, the success of digital archaeology, too, becomes possible only when archaeological informatization is fully realized.

Without informatization, digital technologies can only serve as piecemeal solutions, potentially undermining not only the contextual integrity of archaeology but also the precision and efficiency of various digital operations. For digital archaeology to achieve a substantive transformation, informatization must be both its starting point and its prerequisite.

<sup>8</sup> The variability of archaeological data is not always dramatic; in many cases, new findings can be flexibly accommodated within existing classification frameworks. However, from the standpoint of informatization, the complexity of the problem cannot simply be assessed in terms of “disciplinary accommodability.” Database and programming structures operate according to predefined schemas and input formats; each time a new type of data that falls outside the standard emerges, the entire structure must be modified or exception-handling routines must be repeatedly designed. As these exceptions accumulate, they increasingly burden the consistency and maintainability of the informatization system, raising the likelihood that the initially designed “standard” will quickly become outdated.

## 2. A Review of Existing Discussions on Archaeological Informatization in Korea

The need to organize and systematize archaeological data using digital technologies has been consistently raised. In Korea, related discussions have taken place across multiple disciplines, including archaeology, records management, and digital humanities, and have been presented in various forms—not only as academic papers but also as policy reports and conference presentations. Given that systematic research in this area remains insufficient, this paper draws on the full range of these sources to trace the trajectory of discourse on archaeological informatization in Korea.

Ji-In Kim (2000) called for the construction of a comprehensive database for artifacts and the broader field of archaeological research, proposing both relational and object-oriented database models.<sup>9</sup> Kim further argued that training archaeologists with sufficient knowledge of computer science would be necessary to actively leverage future IT environments.

The National Computerization Agency (2000) assessed that the level of informatization across the cultural heritage sector at the time was generally very low; apart from the National Museum of Korea and the Cultural Heritage Administration, most institutions lacked even a basic website. The report identified centralized governance structures, a shortage of specialized personnel and dedicated organizational units, insufficient budgets, and closed information management systems as the principal causes. It further observed that the divergent management frameworks employed by individual institutions rendered the integrated application of existing software impractical. The agency therefore called for the expansion of informatization personnel and organizational capacity, the securing of stable funding, and—given the high degree of domain expertise required—the direct participation of professionals with relevant knowledge.<sup>10</sup>

The Zininzin Academic Team (2005) identified several limitations of excavation reports, including constraints arising from fixed page dimensions, an imbalanced production cost structure that hindered the effective communication of visual information, and inefficiencies in distribution and storage. The team advocated for the digitization of excavation reports under the banner of “buried cultural heritage informatization” (*maejang munhwajae jeongbohwa*),<sup>11</sup> and further proposed the establishment of an integrated information management system for buried cultural heritage.

---

<sup>9</sup> [Translation note] The original footnote addressed a Korean-language terminological issue. At the time of Ji-In Kim’s publication, standard Korean translations of English IT terms had not yet been firmly established. In current Korean usage, “object” is conventionally rendered as “객체” (gaekche) and “entity” as “개체” (gaeche), making “객체지향형” (gaekchejihyanghyeong) the standard Korean equivalent of “object-oriented.” However, Kim adopted the non-standard form “개체지향형” (gaechejihyanghyeong) based on the author’s own conceptual interpretation. Because this was a deliberate terminological choice within Kim’s published work, the present author could not alter it when citing. The original footnote was therefore included to prevent any confusion that might arise from the interchangeable use of these terms. Since both Korean terms translate identically as “object-oriented” in English, this distinction cannot be meaningfully conveyed in translation. (Original footnote 9)

<sup>10</sup> The report in question addressed the level of informatization across the cultural heritage sector as a whole and contained no specific references to buried heritage or archaeological informatization. Nevertheless, it has been cited and summarized here as a representative source for gauging the overall environment and the limitations of informatization efforts at the time.

<sup>11</sup> As of May 17, 2024, the Cultural Heritage Administration was reorganized as the Korea Heritage Service, accompanied by extensive revisions to related legislation and terminology—including the change from “buried cultural heritage” (*매장문화재*) to “buried heritage” (*매장유산*). In this paper, current terminology is used as a general principle; however, when citing past literature or referencing historical institutional contexts, the terms in use at the time have been retained.

Yim (2011) defined excavation records as encompassing “all recorded information produced during the process of excavating buried cultural heritage or archaeological sites—including planning, investigation, and analysis—such as official documents, measured drawings, data, photographs, field journals, excavation reports, and meeting minutes.” Yim criticized the prevailing excavation reports for being disproportionately weighted toward representational information, and proposed a series of remedial measures: the assignment of dedicated records management personnel, the systematic revision of excavation regulations, and the introduction of an integrated management system for excavation records.

The Cultural Heritage Administration (2013), in alignment with the government’s “Government 3.0 (정부 3.0)” initiative, launched the “Cultural Heritage 3.0 (문화유산 3.0)” policy with the primary objectives of organizing and opening up cultural heritage-related big data. Key tasks included the integration of archival resources and the construction of archive and research information databases. Measures specific to archaeological informatization were also stipulated, such as the provision of regulatory spatial data on buried cultural heritage, open access to excavation report full texts, national funding for surface surveys, and the revision of legislation governing underwater cultural heritage.

Hongyeon Kim (2013) observed that although digital records were already in use during excavation, the restriction of final outputs to printed booklet form limited the utilization of information. To address this, Kim proposed a transition to e-book formats,<sup>12</sup> arguing that functional enhancements—such as multimedia integration, improved compatibility, and the removal of publication constraints—could broaden the scope of utilization. Kim further presented a virtuous-cycle cultural heritage service model encompassing the construction of databases from secondary data derived from excavation reports<sup>13</sup> and the development of content derived therefrom.

Cha (2016) noted that although excavation reports had been digitized in PDF format, they remained dispersed across multiple institutions, making access inconvenient and raising the risk that materials could become irretrievable upon the dissolution of issuing organizations. Cha argued that this fragmentation could erode both user convenience and public accessibility, and called for the development of standardized classification indices organized by period and site type, along with unified search standards.

Ahn (2017) pointed out the fundamental difficulty of accessing information on buried cultural heritage in the first place. While acknowledging that the services provided by Zininzin had made a partial contribution, Ahn argued that inherent limitations persisted due to the revenue structure of private enterprises and the difficulties of collaboration between such firms and excavation institutions. Ahn maintained that public bodies—principally

---

<sup>12</sup> The “e-book” referred to here does not mean a PDF-format digital reproduction of a printed volume, but rather an EPUB. EPUB is an electronic publishing standard designed to support features such as adjustable font size, table-of-contents navigation, and full-text search across diverse devices.

<sup>13</sup> Hongyeon Kim classified materials obtained immediately in the field—such as field logs, original measured drawings, digital source photographs, and raw 3D data—as primary data; digitally processed materials such as selected and edited photographs (panoramas, PVR, etc.), fully rendered 3D data, vectorized drawings, and scientific analysis results as secondary data; and excavation reports and similar publications that synthesize and interpret information derived from sites as tertiary data.

the Cultural Heritage Administration and the Association of Research Institutes for Cultural Heritage—should take the lead in promoting information disclosure and utilization, and assessed that the Cultural Heritage Spatial Information Service likewise remained confined to basic map provision, limiting its practical utility.

Rhee (2018) surveyed international archaeological data archive initiatives, notably tDAR<sup>14</sup> in the United States and ADS<sup>15</sup> in the United Kingdom, and emphasized the need to establish a national digital archive for archaeological data that would ensure long-term preservation and facilitate user access. Rhee proposed the development of a policy framework encompassing legislation and institutional arrangements, the establishment of data standards, and web-based access targeting both specialists and the general public. The study additionally highlighted the potential of mobile applications for recording primary data in the field during excavation.

Jung (2020) took issue with the fact that, despite the ongoing digital transformation of the excavation process, outputs remained confined to printed booklets or PDF files. Jung argued that such format-level limitations could compromise the substantive quality of reports, maintaining that a transition to database-driven formats was necessary and that publications should be repositioned in a supplementary role. The study further proposed a database development strategy distinguishing between legacy and newly generated materials, and presented an interface design informed by case studies of Silla-period tumuli.

The Cultural Heritage Administration (2021) issued a mid- to long-term plan for the digital transformation of the cultural heritage sector, structured in three phases: foundation building (2022–2025), service internalization (2026–2030), and service expansion (2031 onward). The initial phase targeted the construction of big data platforms and cloud infrastructure, followed by the introduction of AI capabilities and the strengthening of analytical capacity, with the long-term goal of scaling up advanced services. To this end, the plan outlined a range of initiatives including the development of an integrated archive, a dedicated data center, and training datasets for AI.<sup>16</sup>

Choi (2022) proposed a set of measures to invigorate digital archaeology, including the formulation of archival standard guidelines, the construction of an integrated archive, and the adoption of data quality management frameworks, while underscoring the need for enhanced research support and legislative reform. Choi also discussed the potential for applying diverse AI technologies to archaeology, such as

---

<sup>14</sup> <https://core.tdar.org>.

<sup>15</sup> <https://archaeologydataservice.ac.uk>.

<sup>16</sup> The plan also included some specialized sub-plans related to buried heritage and archaeological data; however, these were primarily oriented toward the utilization phase following informatization—such as predictive modeling of buried heritage distribution and AI-based restoration of pottery. With respect to foundational infrastructure and system construction for informatization itself, the approaches outlined for the buried heritage sector showed little differentiation from those proposed for other sectors.

relationship-oriented data integration grounded in the Semantic Web<sup>17</sup> and ontologies,<sup>18,19</sup> as well as text mining and deep learning. Subsequently, Choi (2023, 2024) turned attention to the transformative potential of AI-driven search and new modes of archaeological data utilization in the wake of ChatGPT.

Chorong Lee (2022) questioned the simplistic information delivery format of the *Dictionary of Korean Archaeology*, and highlighted the need for a new knowledge service capable of enhancing the utility of archaeological information. Chorong Lee (2023) elaborated on this vision by presenting concrete directions and implementation strategies for digital construction that account for permanence, interoperability, and scalability, while emphasizing the potential for AI integration through semantic data and Open APIs.<sup>20</sup> Lee further advocated for the development of an intelligent integrated management platform capable of synthesizing excavation reports, drawings, photographs, and locational data.

Hyeon Kim and H. Kang (2022) analyzed the digital transformation of the *Encyclopedia of Korean Culture* and proposed an ontology-based semantic data model for the *Dictionary of Korean Archaeology*. They emphasized that the rapidly growing volume of data can no longer be managed through human cognition alone, and argued that ambiguous humanities knowledge must be converted into explicit, machine-readable data. The authors further observed that such a semantic archive holds the potential for expansion through AI-driven inferential reasoning.

D. Kang (2022) analyzed archaeological data platform initiatives in Europe, the United States, and Japan, and argued that a comparable platform is needed in Korea. Kang assessed the *Dictionary of Korean Archaeology* as a competitive asset by virtue of its high-quality, collectively generated knowledge base, and proposed a WebGIS platform built upon it. Kang further presented a hybrid model combining Wikipedia-style openness with an Encyclopaedia Britannica-style editorial review system. Subsequently, Kang (2024) expanded the proposal to encompass a metadata system, location-based ontology, linked open data, and a Semantic Web-based WebGIS framework.

B. Kim (2023) argued that the role of archaeological digital archives should shift beyond mere preservation toward enhancing user-centered accessibility and utilization. To this end, Kim called for the establishment of an institutional foundation, the reorganization of management frameworks, the development of

---

<sup>17</sup> The Semantic Web is a technology that structures information on the web so that it can be understood not only by humans but also by computers. Rather than treating content as plain text, it explicitly defines the meanings and relationships of data, enabling machines to automatically connect and interpret them.

<sup>18</sup> An ontology is a method of knowledge representation that structurally defines concepts within a specific domain and the relationships among them. It is primarily used in the Semantic Web and information systems to enable machines to interpret and process data. Through systematic classification and the explicit specification of relationships, ontologies contribute to enhancing the interconnectivity of information.

<sup>19</sup> For example, when the term *cheongja* (청자) appears on the web, the Semantic Web can refer to the knowledge structures defined in an ontology to help a computer understand that it denotes a type of celadon pottery from the Goryeo period and how it is related to specific time periods, typologies, and other artifacts. This enables information processing at the level of context rather than individual words.

<sup>20</sup> An Open API (Application Programming Interface) is a publicly available programming interface that allows external developers and users to retrieve data or access functionality in a standardized format without special authentication or restrictions. It is widely used by public institutions and private services to facilitate information sharing and extend system capabilities, enabling interoperability among diverse systems.

international cooperation mechanisms, and the construction of platforms grounded in public–private partnerships.

Hoon-Hui Kim (2023) observed that despite the adoption of digital technologies, excavation reports remain confined to paper and PDF formats, with limited data standardization. Kim drew attention to accessibility issues with pre-2014 reports and inconsistencies in formatting specifications, and proposed the standardization of data formats and 3D scan data, as well as the construction of a state-led archive. Kim further stressed the need for the archaeological community to broaden its adoption of new technologies.

Chang-hee Lee (2023) noted that although digital recording has been introduced at excavation sites, degradation in resolution and file quality persists at the final report stage. Lee called for the standardization of reports, the systematization of 3D data accumulation, and the construction of an integrated archive for archaeological information, emphasizing that digital archives should be managed centrally under state leadership.

M. Lee (2024) noted that while the importance of cultural heritage utilization has grown, the digital transformation of the buried heritage sector remains inadequate. Lee argued that a digital platform for the automated storage and sharing of the rapidly expanding body of buried heritage materials is necessary, and proposed that dedicated digital teams be established within professional excavation firms, with data linked to the Korea Heritage Service platform. Lee additionally suggested the development of a generative AI–based research support system and an application for exploring artifacts and archaeological sites.

So and Juho Kim (2025) observed that supplementary records generated during excavation—including 3D data, DEMs,<sup>21</sup> and coordinate information—are only partially incorporated into reports and remain dispersed across individual institutions. They proposed that such materials be integrated into a public management framework to ensure stable preservation, and called for the establishment of regional platforms and the revision of legislation to lay the groundwork for unified management and utilization.

### 3. The Trajectory and Limitations of Archaeological Informatization Discourse in Korea

Discussions on archaeological informatization in Korea have accumulated progressively along generational perspectives. Ji-In Kim (2000) was exceptional for the time in proposing database technologies suited to archaeology; the Zininzin Academic Team (2005) charted a practice-oriented direction under the banner of "buried cultural heritage informatization"; and Yim (2011) emphasized the importance of records management, arguing that securing structured records was a prerequisite for informatization.

Subsequent discussions did not depart significantly from these three foundational contributions, merely elaborating on them to varying degrees. Most remained isolated efforts by individual researchers, failing to build on preceding work or translate into policy, so that the early discourse appears to have faded into obscurity.

---

<sup>21</sup> A DEM (Digital Elevation Model) represents the elevation of a terrain surface as digital numerical values. In archaeology, DEMs are used to analyze and visualize the topography of sites and features.

The first turning point in Korean archaeological informatization discourse came with the Cultural Heritage Administration's "Cultural Heritage 3.0" policy in 2013, which is regarded as the first full-fledged policy to explicitly articulate the need for informatization and open sharing. However, it was not a policy centered on buried heritage, and in practice it amounted to little more than making excavation report PDFs available for viewing. Influenced by this initiative, Hongyeon Kim (2013) proposed extending informatization toward broader cultural utilization, but the impact on academia and policy remained limited, as the study had been presented at the undergraduate level.

Subsequent discourse shifted toward advocating for the construction of a national-level archaeological archive, prompted by practical frustrations in the field. Rhee (2018), in particular, is regarded as a pivotal point of departure for internationally informed discussion, having conducted a broad survey of overseas cases—including the ADS in the United Kingdom and tDAR in the United States—and proposed a development plan for a Korean archaeological digital archive on that basis.

The years 2021–2022 marked a second turning point, as both institutional and academic spheres began to mobilize in earnest: the Cultural Heritage Administration announced its digital transformation plan, Choi published a doctoral dissertation on the subject, the National Research Institute of Cultural Heritage hosted an international symposium, and a new Digital Heritage Team was established. With these developments, what had been a discourse driven primarily by individual researchers expanded into concrete proposals for technology adoption and strategic planning.

Yet even after this turning point, tangible outcomes remained scarce. Discussions continued, but most stayed at the level of idealistic discourse, and the technologies proposed rarely went beyond referencing cases from other fields. The number of initiatives actually implemented has been limited to a handful—such as the digitization of the *Dictionary of Korean Archaeology* and GIS applications for urban archaeological sites (National Research Institute of Cultural Heritage 2024: 1–33)<sup>22</sup>—and experiments reflecting the distinctive policy environment and disciplinary character of archaeology have been rare. In the end, the discourse has repeatedly circled back to the same issues of budget, personnel, institutional frameworks, and organizational capacity first identified by the National Computerization Agency in 2000, and it remains unclear whether the Cultural Heritage Administration's (2021) digital transformation plan can be meaningfully realized in the field of archaeology.

#### 4. Structural Constraints and New Possibilities in Archaeological Informatization

Although discussions on archaeological informatization have built a broad consensus over the years, the fundamental reason they have struggled to translate into practice lies not in a lack of initiative on the part of any

---

<sup>22</sup> The National Research Institute of Cultural Heritage maintains and provides open access to various research outputs through its NRICH Archive (국가유산 지식이음, portal.nrich.go.kr). However, this amounts to merely making digitized data searchable, and does not yet appear to have reached the level of informatization discussed in this paper (accessed May 4, 2025).

particular institution or group, but rather in the inherent characteristics and structural constraints of archaeology as a discipline.

Choi pointed out that a gap in perspectives exists between archaeologists and technologists regarding what technological applications are needed in the field (Choi 2022: 34). When technology development is pursued unilaterally without adequately reflecting archaeological requirements, it becomes difficult to achieve the desired outcomes. Chang-hee Lee similarly noted that the Arch3D Liner program, developed by carrotphant Inc., was able to incorporate archaeologist-friendly features precisely because its founders were trained in archaeology (Chang-hee Lee 2023: 162–164). This demonstrates that adequate system design is possible only when the complex interplay between archaeological data and technology is fully understood.

B. Kim explained that because archaeology is a discipline that deals with the vast temporal span of human society and culture, the types and forms of data it handles are inevitably diverse and extensive (B. Kim 2023: 6). Takata and Ho further noted the difficulty of providing multilingual access to archaeological materials and integrating platforms across cultural and linguistic barriers, suggesting that archaeological data possess characteristics that resist capture within a single standard or a simplified system (Takata and Ho 2022: 38–42).

Moreover, the pursuit of IT projects by government and public institutions is subject to structural constraints. As the volume and complexity of data grow, the design capabilities of institutions themselves reach their limits, and external contractors also struggle to adequately reflect the archaeological context. Most contractors operate solely within the scope of a Request for Proposal (RFP), with limited capacity for flexible modifications or accommodation of disciplinary specificities. Constraints of time and budget, compounded by complex administrative procedures, further exacerbate these issues—a problem widely recognized not only in archaeology but across the IT outsourcing sector as a whole.<sup>23</sup>

It is against this backdrop of structural constraints that LLMs have recently emerged as a technology offering new possibilities for archaeological informatization. Whereas conventional approaches to informatization presupposed highly formalized system design, extensive predefined schemas, and the sustained involvement of experts versed in archaeological context, LLMs—as language-based artificial intelligence trained on massive natural-language corpora—possess the potential to partially replace or dramatically streamline these processes. Their capacity to automate tasks such as classification, summarization, structuring, and question-answering, and to flexibly reorganize information in context-sensitive ways, makes them a potentially invaluable tool for a field as unstructured and demanding of specialized expertise as archaeology.

---

<sup>23</sup> The conventional IT project execution framework carries not only the problems already apparent but also more fundamental limitations that will manifest before long. The digital technology landscape, LLMs included, has recently been evolving at a pace incomparable to the past. GPT and Gemini release new versions two to three times per year, and Claude three to four times per year; considering these major LLMs alone, a new model emerges roughly every one to two months. When open-source models are factored in, the sheer volume becomes far greater still. Against this pace of technological change, the conventional IT project execution framework—with its protracted cycles from planning to implementation—struggles to keep pace. A fundamental restructuring of this framework is therefore necessary going forward.

Choi (2023, 2024) and M. Lee (2024) discussed the potential for introducing generative AI, including ChatGPT, and proposed approaches for its application in the cultural heritage and buried heritage sectors. Their focus has primarily been on end-user applications such as chatbots and research assistance tools. I share their view that generative AI should serve as a practical tool for both researchers and the general public, and this paper addresses some of those possibilities as well. However, the present study places greater emphasis on constructing the informatization infrastructure that underpins such applications, with particular attention to the role of LLM technology as a means of realizing that goal.

### III. The Concept of LLM Technology and Its Potential for Archaeological Informatization

#### 1. The Concept of LLMs and Their Differences from Conventional AI

In archaeology, artificial intelligence (AI) based on machine learning and deep learning has already been employed for tasks such as image-based artifact classification and pattern recognition in structured attribute data. However, LLMs, which have attracted growing attention in recent years, represent a new paradigm that demonstrates a qualitatively different level of performance in language-based data processing compared to conventional AI.

LLMs are deep-learning-based artificial intelligence trained to understand and generate natural human language. What distinguishes them from earlier approaches is their sheer scale: they are models comprising a vast number of parameters<sup>24</sup>—ranging from tens of billions to trillions (Zhao, W. X. et al. 2023). Through this massive architecture and training, LLMs have acquired high-level linguistic capabilities that go beyond simple statistical pattern analysis—including grammatical knowledge, factual information, reasoning, and contextual understanding.

The core technological foundation underlying this performance is the Transformer architecture, proposed in 2017. Earlier models such as Recurrent Neural Networks (RNNs)<sup>25</sup> and Long Short-Term Memory models (LSTMs)<sup>26</sup> required sequential data processing, while Convolutional Neural Networks (CNNs)<sup>27</sup> were primarily specialized for image analysis. The Transformer, by contrast, was designed to simultaneously compute relationships among all words within a sentence and assess their contextual importance through a Self-Attention Mechanism.<sup>28</sup> This enables effective comprehension even in long sentences or complex contexts, and the architecture is regarded as particularly well suited for handling long-range dependencies in language (Vaswani et al. 2017).

The training process of LLMs also differs fundamentally from that of conventional AI models. In the pre-training stage, self-supervised learning is performed on vast volumes of unstructured text data, enabling the model to acquire language patterns, grammar, and general knowledge without the need for explicit labels (Zhao, W. X. et al. 2023; Brown, T. B. et al. 2020). This represents a departure from the supervised learning approach

---

<sup>24</sup> Parameters are numerical values that an AI model adjusts through training; they are the factors that determine the relationship between inputs and outputs. A higher parameter count enables the model to learn more complex patterns, but proportionally greater computational resources and training data are required.

<sup>25</sup> A type of AI model used to process sequential data such as sentences. It predicts the next word while retaining information from preceding words, but its ability to retain earlier context diminishes as sentences grow longer.

<sup>26</sup> A variant of RNN designed to selectively retain important information over extended sequences while promptly discarding less relevant information. This architecture enables the model to preserve earlier context relatively well, even in long sentences or complex contexts.

<sup>27</sup> A type of AI model designed to effectively process data with spatial structure, such as image and video data. It excels at automatically extracting features within photographs and is primarily used for tasks such as classifying artifact photographs or analyzing satellite imagery.

<sup>28</sup> A mechanism that simultaneously evaluates the relationships among all words in a sentence, enabling the model to autonomously determine which words are important. For example, in the sentence “He ate the apple,” it helps the model identify from context whom “he” refers to. This mechanism allows the model to effectively comprehend even long sentences or complex contexts.

of conventional models, which rely on structured, labeled data. In a subsequent fine-tuning stage, the model is adapted to specific tasks or specialized domains—such as archaeology, with its distinctive terminology and narrative structures—while retaining its broader linguistic capabilities. This process employs not only supervised learning but also Reinforcement Learning from Human Feedback (RLHF),<sup>29</sup> in which human evaluations are used to further refine the model's outputs (Ouyang, L. et al. 2022).

These characteristics of LLMs have the potential to create a significant turning point for archaeological informatization. Conventional machine learning and deep learning models have primarily served as task-specific tools—for example, CNN-based artifact photograph classification, satellite imagery analysis for site prospection, and analysis of structured excavation data. LLMs, by contrast, are capable of flexibly analyzing vast quantities of unstructured textual materials such as excavation reports, research papers, and artifact descriptions; extracting or summarizing specific information from them; and synthesizing relevant data to respond comprehensively to complex research questions. In particular, for materials with recurring formats and content structures—such as excavation reports—LLMs demonstrate considerable potential for automatically converting them into structured data.

This capacity for automation and information reorganization serves not only to maximize the efficiency of archaeological informatization, but also to form a foundation for enhancing knowledge accessibility and strengthening both the precision and scalability of research.

## 2. Use Cases of LLMs

### 1) General Use Cases

In recent years, commercial services based on various LLMs—such as ChatGPT, Claude, Gemini, and Perplexity<sup>30</sup>—have proliferated rapidly, transforming the way a wide range of language-based tasks are performed, from everyday communication to the application of specialized knowledge. These tools go beyond simple information retrieval or autocomplete functions: they carry out complex, multifaceted operations encompassing language generation, comprehension, summarization, classification, and reasoning, functioning as digital assistants closely aligned with human thought processes.

LLMs have found their most active application in the field of programming. Tools such as GitHub Copilot<sup>31</sup> enhance developer productivity through automated code completion, function documentation, and

---

<sup>29</sup> A training method in which human evaluators distinguish between preferred and suboptimal responses, teaching the AI which outputs are more desirable. Just as one teaches a child which expressions are polite, this approach trains the AI to produce more natural and appropriate responses through human feedback.

<sup>30</sup> The four services mentioned above represent relatively popular and general-purpose use cases. Beyond these, commercially available LLM-based services specializing in specific functions—such as editing video via voice commands or automatically generating presentation materials—are rapidly proliferating across various fields. Those identified by the author alone number in the dozens, and tools that simultaneously integrate domain expertise and automation continue to be released in areas ranging from education and law to marketing and design. This trend suggests that LLM technology is establishing itself not merely as a text generation tool but as a core instrument for automating and advancing a wide range of real-world tasks.

<sup>31</sup> <https://github.com/features/copilot>

generation of repetitive structures. More recently, a practice known as “vibe coding” has gained traction,<sup>32</sup> in which code is progressively built through an iterative exchange of natural-language instructions and AI responses; this approach has attracted attention as particularly well suited for early-stage prototyping and creative problem-solving. LLMs now function not merely as automation tools but as assistants that interpret developers’ intentions and provide real-time support.

In this way, LLMs have moved beyond serving as simple language assistance tools to become deeply integrated into a wide range of knowledge-based tasks, transforming the very manner in which work is conducted. Across diverse domains—including document composition, information summarization, knowledge exploration, and software development—LLMs have already established themselves as practical, real-world tools.

## 2) LLM-Based Informatization Cases

In recent years, LLM-based technologies have been applied to the structuring of unstructured text across diverse fields, including medicine, the social sciences, and journalism and information studies. Capabilities now extend beyond simple summarization to include key-value extraction and structuring into formats such as JSON.<sup>33</sup> A growing body of cases demonstrates that major models—including GPT, Claude, and LLaMA—can achieve high levels of accuracy through sophisticated prompt design alone.

In the medical field, active efforts are underway to use LLMs for structuring unstructured clinical data such as electronic health records (EHRs) and pathology reports. A 2024 study found that models including GPT-4 and Claude 3 extracted key fields—such as patient ID, admission date, and risk scores—with an accuracy exceeding 0.98, significantly outperforming earlier RoBERTa-based models.<sup>34</sup> In a separate analysis of pathology data, ChatGPT extracted pathological grades, resection status, and similar items with over 90% accuracy using fixed prompts alone, surpassing conventional NLP techniques<sup>35</sup> without any task-specific retraining (Ntinopoulos et al. 2025). These findings suggest that LLMs can be effectively applied to the automatic extraction and structuring of standardized key information from archaeological excavation reports—such as feature numbers, artifact names, stratigraphic layers, and measurements.

In the field of qualitative data analysis, LLMs are being actively used to analyze unstructured texts—such as interview transcripts and counseling conversations—and derive thematic or coding frameworks from them.

---

<sup>32</sup> “The Crossroads Developers Face: Learn Vibe Coding, or Retire” (개발자가 맞닥뜨린 갈림길 ‘바이브 코딩’을 배우거나, 은퇴하거나). ITWorld Korea. Accessed May 4, 2025, <https://www.itworld.co.kr/article/3967678>.

<sup>33</sup> JSON (JavaScript Object Notation) is a lightweight text format for representing data structures and objects, composed of key-value pairs. It is easily readable and writable by both humans and machines and is widely used in web environments and APIs. For example, a single artifact record expressed in JSON would take the form: {"artifact": "bronze mirror", "period": "Three Kingdoms period", "site": "Gyeongju"}.

<sup>34</sup> RoBERTa (Robustly Optimized BERT Approach) is a natural language processing model developed by Facebook AI in 2019 as an improved version of BERT (Bidirectional Encoder Representations from Transformers). It is characterized by substantially enhanced performance achieved through larger training datasets, longer training duration, and modifications to the masking strategy, demonstrating high accuracy across a wide range of natural language understanding tasks.

<sup>35</sup> NLP (Natural Language Processing) refers to the broad range of techniques that enable computers to understand and process human language. Traditional methods include morphological analysis, part-of-speech tagging, syntactic parsing, named entity recognition (NER), sentiment analysis, and text classification. In recent years, deep learning-based models have been replacing or supplementing these methods, thereby driving further advancement.

For example, in a study analyzing interviews with parents of children with rare cardiac conditions, GPT-4-series models extracted themes at a level comparable to that of human analysts, achieving both efficiency and quality through a LangChain-based pipeline<sup>36</sup> and strategic chunking techniques.<sup>37</sup> In analyses of interviews with medical professionals, methods were introduced for summarizing responses by question and automatically clustering themes; in the classification of counseling strategies and the extraction of dynamic indicators from therapeutic dialogues, GPT-4 likewise achieved the highest accuracy, approaching the level of human experts (Raza et al. 2025). Such approaches could contribute to automatically classifying and summarizing specific research topics, interpretive perspectives, and period-specific characteristics from the vast body of excavation reports and research literature, thereby facilitating the identification of research trends and comparative analysis.

Large-scale information sources such as academic literature and news articles are also being structured through LLMs. In the summarization of scientific papers, studies have emerged that replace the conventional complex pipeline of named entity recognition (NER) and relation extraction<sup>38</sup> with prompt-based JSON output approaches; Meta's LLaMA-2 and OpenAI's GPT-series models have successfully structured experimental conditions, material properties, and attribute relationships with high efficiency. In the domain of news articles, one study used ChatGPT to automatically extract and structure information about protest events—including dates, locations, organization names, and tactics—with accuracy in some fields exceeding that of manual coding (Caren et al. 2024). These cases similarly demonstrate that LLMs can be applied to the consistent extraction and database construction of information from large numbers of archaeological reports and papers—such as survey methodologies for specific sites, typological classification criteria for artifacts, and the bases for chronological dating.

These cases demonstrate that LLM-based structuring of unstructured text is yielding tangible gains in operational efficiency and the conversion of information into data assets. Challenges remain, however, including gaps in domain-specific knowledge, sensitivity to prompt design, and limitations in contextual processing. To address these issues, hybrid approaches that combine LLM outputs with expert review have recently attracted growing attention.

The LLM-based informatization cases validated in fields such as medicine and the social sciences offer significant implications for archaeology as well. LLMs show particular strength in the automatic structuring and

---

<sup>36</sup> LangChain is a framework that facilitates the development of complex language applications by connecting large language models (LLMs) to various external data sources (documents, databases, etc.) and tools. It provides advanced capabilities such as prompt chaining, memory management, agent configuration, and tool integration, and is widely used for implementing LLM-centric automation pipelines.

<sup>37</sup> Chunking is a preprocessing method that divides text into manageable units so that an LLM can process them, splitting long documents into multiple segments accounting for the model's context window limits. Strategic chunking refers to an approach that, rather than splitting solely by length, divides text along semantic units (paragraphs, topics, conversational turns, etc.) to preserve information and minimize contextual loss. This technique is particularly effective for improving LLM performance in the analysis of unstructured text such as lengthy interviews or counseling transcripts.

<sup>38</sup> NER (Named Entity Recognition) is a natural language processing technique that identifies proper nouns or specific entities such as people, locations, dates, and organizations within a sentence. Relation Extraction refers to the process of determining the relationships among these recognized entities, generating information structures such as "Person A was active at Place B." The combination of these two steps is widely employed in traditional metadata extraction pipelines and has served as a principal approach for constructing structured data from unstructured text.

summarization of excavation reports, which possess consistent narrative patterns, and they point toward a new direction for informatization in that they enable flexible access to information through natural-language queries without requiring the prior integration of data in disparate formats.

### 3. Rewriting the Strategy for Archaeological Informatization with LLMs

Conventional approaches to archaeological informatization have primarily been grounded in established technological frameworks such as the Semantic Web, ontologies, and data mining. Each of these has evolved with distinct objectives—structuring information, defining relationships among concepts, discovering patterns, and enabling semantics-based analysis—and represents the cumulative product of years of research and standardization efforts. However, these approaches fundamentally presuppose extensive upfront design, metadata definition, relationship specification, and inter-institutional collaboration, all of which demand considerable time and resources in their own right. Given the complexity of archaeological data—with its diverse sources and formats—and the prospect of continuously accumulating data in the future, the standardization work required for consistent structuring may be realistically unsustainable.

Informatization based on the Semantic Web and ontologies has pursued long-term integration primarily in Europe, the United Kingdom, and Japan, drawing on international standards such as CIDOC CRM.<sup>39</sup> While such approaches are advantageous for semantically consistent information management, they require a high level of upfront design, coordination, and funding, and there is a considerable gap between these requirements and the realities of the Korean archaeological information environment. In particular, it is difficult to organize the vast body of unstructured reports that has already accumulated using conventional methods alone.

By contrast, the recent emergence of LLMs opens the possibility of a fundamental shift in the very approach to informatization. LLMs can directly comprehend and process unstructured text without the prior definition of structures or standards, recognizing recurring descriptive conventions and automatically structuring the information they contain. Their ability to follow the semantic flow and logical connections of sentences and to draw inferences from context makes them particularly well suited for identifying and organizing the units of information scattered throughout archaeological narratives. For example, an LLM that has learned the repetitive formats and descriptive patterns of excavation reports can automatically extract and summarize artifact quantities, locations, stratigraphic layers, and typological categories, and transform this information into a structured database.

Excavation reports, in particular, are ostensibly unstructured narrative data, yet they possess semi-structured characteristics owing to their recurring descriptive conventions and consistent use of terminology. LLMs are a technology remarkably well aligned with these characteristics: without separate preprocessing or tagging, they can autonomously organize artifact attributes, excavation contexts, and

---

<sup>39</sup> CIDOC CRM (Conceptual Reference Model) is a conceptual reference model for the semantic integration of cultural heritage information, developed by the Documentation Committee (CIDOC) of the International Council of Museums (ICOM). It aims to describe heterogeneous cultural heritage information—such as museum artifacts and archaeological data—within a common ontological framework and to ensure interoperability. It has been adopted as a core standard in numerous digital cultural heritage projects across the European Union.

comparative references. Their ability to extract meaningful information from within documents and reconstitute it in a consistent structure holds genuine potential to replace existing structuring frameworks.

Moreover, even when report formats or metadata standards differ across institutions, LLMs possess the flexibility to identify similar information based on context and integrate it for analysis. As Jung (2020: 21) and Choi (2022: 32–33) have noted, practical constraints arise from the fact that report formats, data specifications, and metadata standards vary from institution to institution, making integration difficult. LLMs, however, have the capacity to autonomously reconstitute such heterogeneous data into a unified format or automatically convert it to conform to a designated schema.

The prevailing assumption has long been that archaeological informatization can proceed only when various preconditions—institutional frameworks, budgets, and personnel—are in place. Today, however, LLM-based pipelines can be implemented without large-scale infrastructure, prompting a reconsideration of the conventional premise of “informatization only after full preparation.” A pragmatic approach that combines experimentation and construction around currently available technologies may, in fact, prove more realistic.

Interestingly, the emergence of these technologies suggests that not having forced informatization through conventional means may, in retrospect, have been the better choice. Traditional approaches required complex preparatory steps such as establishing standards and designing systems, whereas LLM-based informatization has made a certain degree of structuring and utilization possible without such prerequisites. The technology continues to evolve, and more sophisticated information interpretation and relational inference can be anticipated in the future.

This trajectory also resonates with the distinctive nature of archaeology itself. Just as excavation entails destruction, the timing and method of informatization matter critically. Rather than forcing structuring when the technology is not yet mature, pursuing it at a turning point such as the present offers greater advantages in both quality and efficiency. Archaeological informatization must now be redesigned around a new, technology-driven strategy rather than repeating the approaches of the past.

To be sure, the LLM-based approach is not intended to replace existing informatization methods altogether. Ontologies, the Semantic Web, GIS-based integrated systems, and similar frameworks continue to play important roles in information structuring and retrieval. What has changed, however, is that it is now more realistic and efficient to position LLMs as the central technology rather than a mere auxiliary tool, combining them with conventional methods as needed. LLMs open new possibilities—contextual understanding, automatic structuring, and question-answering—and are emerging as a key technology that can both complement and accelerate established approaches to informatization.

#### 4. Post-Informatization Applications of LLMs<sup>40</sup>

---

<sup>40</sup> This section may digress slightly from the paper’s central discussion of archaeological informatization; however, it briefly presents practical use cases in order to illustrate the extensibility and versatility of LLM technology.

LLMs also effectively expand the scope of utilization of constructed databases in the post-informatization phase. This expansion can be understood in terms of two roles: assisting researchers in factual information retrieval and enhancing public accessibility to knowledge. For researchers, LLMs function as research assistance tools that rapidly extract and synthesize specific factual information from vast corpora, thereby freeing up time for deeper analysis. For the general public and cultural content professionals, they can lower the threshold to archaeology by providing accessible summaries to complex questions—such as “What shapes of earrings did people wear during the Silla period? Where were such artifacts mainly found?”—or offer inspiration for new content creation. It must be noted, however, that directly incorporating the “interpretations” produced by LLMs themselves into professional research carries the risk of factual distortion and must be guarded against at present; this is an issue that can only be meaningfully discussed in a distant future when the technology has attained an extremely high level of maturity.

Furthermore, LLMs can comprehend and generate text in multiple languages, producing consistent translation outputs even if they do not yet reach professional-level quality. In particular, integrating the English edition of the *Dictionary of Korean Archaeology* via RAG<sup>41</sup> enables more precise translations grounded in prior knowledge of archaeological terminology. When combined with expert post-editing of AI-generated translations, this can serve as a pragmatic strategy for reducing the cost and time required to develop foreign-language services.

Meanwhile, it has long been noted that archaeological data across nations are difficult to link due to differences in language, culture, and format. LLMs, however, can contextually comprehend diverse expressions and infer analogous concepts, thereby enabling at least a provisional integration of heterogeneous information. Even before international standards are fully established, the potential to serve as a bridge that meaningfully connects data worldwide is significant.

Moreover, beyond excavation reports, LLMs can automatically link a wide range of specialized sources—archaeological papers, conference presentations, the *Encyclopedia of Korean Culture*—as well as non-specialist content available on the web, such as news articles, blogs, and YouTube descriptions. This represents a new approach in that connections between information can be established on the basis of context, without the manual design of relationships required by the Semantic Web or ontologies. Publicly available materials for which crawling is permitted can be utilized without separate agreements and integrated with archaeological information without additional processing.

## 5. Limitations of LLMs and Strategies for Mitigation

As with all technologies, LLMs possess inherent and technical limitations, and their practical application is feasible only when accompanied by strategies that clearly identify and adequately address these constraints.

---

<sup>41</sup> The concept of RAG is addressed in the subsequent section.

The representative limitations of LLMs can be summarized in three categories. First, AI hallucination (factual distortion) is among the most widely recognized issues. This refers to cases in which the model confidently generates incorrect information or presents content that does not exist in the source documents as though it were factual. Second, there is the limitation of temporal currency. Most commercial LLMs cannot reflect information beyond their training cutoff and frequently lack up-to-date information or domain-specific knowledge. Third, the absence of idempotency emerges as a significant concern in practical applications. Idempotency refers to the property of guaranteeing identical outputs for identical inputs; current LLMs, however, tend to return slightly different responses each time even when the same text is entered repeatedly.

Various strategies to address these limitations are also advancing in tandem. A representative example is the RAG (Retrieval-Augmented Generation) architecture, in which an LLM retrieves relevant information from external documents or databases and then generates responses on that basis. This enables the simultaneous resolution of two challenges: ensuring temporal currency and strengthening factual accuracy. Protocols such as MCP (Model Context Protocol)<sup>42</sup> are also attracting attention as a similar approach that can mitigate hallucination and temporal currency issues through integration with external information sources.

In addition, prompt engineering—the technique of strategically designing input statements to improve the quality of a model’s responses—has recently seen active efforts to automate the process or systematize it into domain-specific templates. Alongside this, domain-specific fine-tuning is a strategy that involves additionally training a model on specialized documents from a particular field in order to achieve greater precision and domain expertise.

Finally, the Expert-in-the-Loop (EITL) model is an approach in which, rather than having the LLM perform all judgments independently, human experts collaborate with it to correct errors and ensure reliability. Particularly in fields with a low tolerance for error, this can be effectively employed as a hybrid operational model in which practitioners and LLMs share roles, jointly carrying out information structuring and analysis.

What is noteworthy is that these mitigation strategies do not merely resolve the shortcomings of LLMs but also generate new currents in informatization. For instance, a workflow in which an LLM summarizes a report, automatically generates key queries from that summary, and then extracts answers from the original text to present to the user can function as an integrated informatization pipeline that transcends a simple question-and-answer system. This ultimately demonstrates the potential to establish a new model of digital transformation in which diverse tasks—data structuring, question-answering, automatic metadata extraction, and document summarization—are unified into a single continuous workflow.

---

<sup>42</sup> MCP (Model Context Protocol) is an open standard introduced by Anthropic in 2024 that enables language models to operate with greater sophistication by incorporating diverse contextual information—such as prior conversations, external documents, and application states—beyond simple question-answering. For example, when composing an email with the assistance of an LLM, a standalone commercial LLM application such as ChatGPT can only generate the content in response to a prompt, after which the user must manually copy it into a mail client and send it. However, when a mail application such as Gmail is integrated with an LLM via MCP, the interaction extends beyond mere composition: a unified command such as “Write an email to X with this content and send it at 3 PM” becomes possible, encompassing composition, delivery, and scheduling in a single operation.

Thus, to successfully apply LLMs to archaeological informatization, it is essential to clearly recognize their limitations and actively employ the mitigation strategies discussed above—RAG, prompt engineering, domain-specific fine-tuning, and expert review. Given the distinctive nature of archaeological data and the rigorous standards of scholarly accuracy, a hybrid approach that combines the automated processing capabilities of LLMs with the deep knowledge and critical scrutiny of human experts is indispensable.

## IV. A Proof of Concept for LLM-Based Archaeological Informatization

### 1. A Preliminary Experiment Using a Commercial LLM Client

Anyone with experience using commercial LLM clients such as ChatGPT may have wondered: “Would it not suffice to simply attach an excavation report PDF and write a well-crafted prompt? As time passes, will the model not become more sophisticated, naturally yielding better results?” To verify what such an approach actually produces, I conducted a simple experiment.

#### 1) Process

When I conducted an informatization experiment based on excavation reports using commercial LLM clients, file size limitations rendered Claude (30 MB) and Gemini (100 MB) practically unusable.<sup>43</sup> Since the majority of excavation reports currently exceed 100 MB, ChatGPT was effectively the only viable tool. Accordingly, I carried out the experiment using the GPT-4o model available through ChatGPT Plus, targeting the report *The Baekje Garden and Trail Site in Hwajisan Mountain, Buyeo*.<sup>44</sup>

Through the prompt, I requested not a simple summary but the extraction of structured data centered on features and artifacts. Features and artifacts were each to be treated as individual entities, and the output was required in CSV<sup>45</sup> format including detailed fields such as measurements, drawing numbers, stratigraphic layers of recovery, coordinates, and artifact–feature relationships. Each query took an average of approximately six minutes to receive a response, and the results were mostly limited to a general level, necessitating repeated follow-up prompts. Although a CSV file was generated, a system error prevented its direct download; instead, the results were rendered as tables form within the chat interface. Even this output covered only a subset of items at a time, and attempting to output the entire dataset in a single pass resulted in excessively long processing times, confirming the practical limitations of this approach.<sup>46</sup>

#### 2) Evaluation

The approach of directly uploading excavation reports and extracting information through commercial chatbot-based LLM clients has the advantage of accessibility in that documents can be analyzed without any development effort. While useful for simple summaries or queries, this method revealed practical limitations with large-scale documents: processing times were lengthy, results were unstable, and repeated requests again consumed considerable time. In particular, sophisticated linkage with images, consistent structuring, and batch processing of multiple documents proved difficult to achieve through a chatbot interface alone.

Above all, converting structured results into persistently usable information by storing them in a database requires guaranteed schema consistency and accurate field matching—conditions that are difficult to fulfill

<sup>43</sup> Verified on May 8, 2025.

<sup>44</sup> Conducted on May 9, 2025.

<sup>45</sup> CSV (Comma-Separated Values) is a text-based file format in which data fields are delimited by commas, providing a simple and intuitive structure for storing and exchanging tabular data. It is compatible with most data processing software, including spreadsheet applications such as Excel, and is widely used for database construction and structured data management.

<sup>46</sup> Due to space constraints, the detailed process and results of this experiment could not be fully included in the main text and are instead provided as an image. For further details, refer to <https://poc.heripo.com/chatgpt-experiment-result.png>.

reliably through a chatbot-based approach. While one-off information extraction may be feasible, pursuing the construction of structured data encounters structural constraints in terms of both performance and consistency.

Accordingly, a pipeline approach that combines the natural language processing capabilities of LLMs with traditional preprocessing tools is necessary for full-scale informatization work. A hybrid strategy—employing rule-based methods for structurally well-defined areas and LLMs for areas requiring interpretation—is considered the most realistic alternative in terms of repeatability, precision, and automation.

## 2. Proof of Concept for an Archaeological Informatization Pipeline

In the preceding section, the experiment using a commercial LLM client confirmed that a chat-based interface alone has inherent limitations for structuring and informatizing archaeological reports. This section accordingly constructs a custom pipeline combining LLMs with conventional programming tools and carries out a proof of concept, in order to realize a more precise and repeatable informatization process.

In this process, I did not use a general-purpose LLM framework such as LangChain but instead designed and implemented the pipeline directly. While LangChain offers advantages in ease of configuration, its generalized architecture can become a constraint when dealing with data of high specificity and irregularity such as excavation reports. Drawing on my domain expertise as an archaeology specialist and my technical capabilities as a software engineer, I determined that it would be more effective to build a flexible, purpose-specific architecture from scratch.

The overall system consists of four modules, each performing the following role: (1) the PDF processing module extracts text and images from reports; (2) the automatic metadata extraction and structuring module uses LLMs to extract and formalize information on sites, features, and artifacts from the text; (3) the database module stores and manages the structured information in an RDBMS<sup>47</sup>; and (4) the web interface module provides PDF upload, result visualization, and browsing functionalities.<sup>48</sup> [Translation note: The original text erroneously states “five modules” and numbers them (1), (2), (4), and (5), omitting (3). The actual initial design comprised four modules, and the translation corrects both the count and the numbering accordingly. Later in the paper, two additional modules are introduced as the system evolved during development, bringing the total to six; this paragraph describes the original four-module design.]

This configuration is focused not on intervening manually at the level of archaeological interpretive depth or fine-grained attribute detail, but rather on verifying the operational feasibility of the technical concept. In other words, the primary objective is to examine whether it is actually possible to achieve the effect of “placing an excavation report into a magic box and having a database generated on the spot.” If the structured output

---

<sup>47</sup> RDBMS (Relational Database Management System) is a relational database management system that organizes data into tables, enabling the definition and management of relationships among data items. In archaeological informatization, it provides a structure for clearly defining and querying the relationships among entities such as sites, features, and artifacts.

<sup>48</sup> The specific technology stacks used in each module are not individually enumerated in the main text; this reflects the author’s personal choices made to prioritize efficiency and speed during the proof-of-concept phase. As the intent is not to prescribe an optimal solution for the informatization pipeline, the combination of technologies may be readily substituted depending on the actual implementation environment or objectives.

attains a certain threshold of completeness, then expanding or fine-tuning the attributes for individual sites, features, and artifacts can be carried out with relative ease thereafter. Even if initial accuracy is somewhat low, the fact that programming logic can be refined and prompts can be made more sophisticated to incrementally raise accuracy once the concept has been proven lends this approach sufficient significance as a starting point for evaluating technical feasibility.

At present, LLMs are being employed at the stage of information extraction and structuring, but this foundation can subsequently be extended to a variety of downstream applications such as visualization (e.g., map-based feature distribution), specific question-answering, and comparative analysis of artifacts.<sup>49</sup> This section presents the process of proof of concept, focusing on the pipeline implementation and initial results, as a point of departure for these possibilities.

This proof of concept was developed and executed on a MacBook Pro 14-inch (2021, Apple M1 Pro, 32 GB RAM) running macOS 15.4.1.

It should be added that by the time this paper is published, a far more advanced structuring system than the results presented in this PoC will in all likelihood already be in place. Generative AI—and LLM technology in particular—is evolving at a remarkable pace, and I intend to continue refining and enhancing the system beyond the scope of this study. This technological evolution will persist well past the point of publication, and the digital appendix webpage will be continuously updated with improved systems and results as they are implemented. Readers wishing to examine the most current implementations and performance firsthand are encouraged to visit the webpage. [Translation note: As of April 2026, when this translation was prepared, a substantially more advanced open-source successor project, heripo engine (<https://github.com/heripo-lab/heripo-engine>), is actively under development. The project is also listed on Open Archaeo (<https://open-archaeo.info/post/heripo-engine>), an international directory of open-source archaeological software.]

## 1) Process

### (1) PDF Processing Module

The primary objective of the PDF processing stage was to effectively separate and extract text and images from excavation reports using appropriate libraries in a Python environment. Although similar tasks can be performed through LLMs, text and image extraction is already a mature technological domain, and a Python-based approach was deemed more appropriate in terms of consistency and cost efficiency.<sup>50</sup>

Initially, five reports were selected for testing, but two of these were vector-based image PDFs from which text extraction was impossible without OCR (optical character recognition). Given that this was a proof of

---

<sup>49</sup> As a result, certain question-answering and artifact comparative analysis functions were partially implemented in this PoC; however, these were not included in the initial planning stage but emerged as an extension during development, as the potential applications of the structured data were explored.

<sup>50</sup> pdfplumber was employed—a Python library designed for more precise extraction of diverse elements such as text, tables, and images from PDF documents. It is particularly advantageous for sophisticated data extraction that accounts for visual structure, as it provides positional information (bounding boxes) for text within each page.

concept, incorporating OCR would have expanded the scope excessively, and these documents were therefore excluded from the analysis. The final selection comprised three reports: *The Baekje Garden and Trail Site in Hwajisan Mountain, Buyeo* (hereafter the Hwajisan report), *Jeju Hangpaduri Hangmong Antimeral Resistance Site (7th)* (hereafter the Hangpaduri report) [Translation note: “Antimeral” in the report’s published English title is not a standard English word; the intended meaning is “Anti-Mongol” (抗蒙), referring to the Goryeo dynasty’s resistance against the Mongol invasions. The translation retains the published title as-is.], and *Seokjangri Paleolithic Site (14th), Gongju* (hereafter the Seokjangri report).

The secondary objective—automatic matching between images and their captions—proved difficult to implement precisely with Python alone, owing to the variation in layout and descriptive conventions across reports. Accordingly, positional coordinates were stored alongside each image and text element, so that in a subsequent stage the LLM could assist in establishing automatic linkages on the basis of contextual information.

The time required for text and image extraction was approximately 88, 29, and 138 seconds for the Hwajisan, Hangpaduri, and Seokjangri reports, respectively.

## (2) Caption Matching Module

This module was not included in the original plan but was added after its necessity became apparent during development. It was initially intended to be implemented within the PDF processing stage; however, since determining the semantic association between images and captions proved difficult through traditional rule-based methods alone, it was separated out as a dedicated LLM-based module.

Specifically, a prompt was designed to instruct the LLM to identify the caption corresponding to each image based on the distance between the coordinates of each text line extracted by the PDF processing module and the coordinates of the image. Since feeding an entire report spanning hundreds of pages into the LLM each time would incur substantial cost, the initial test was limited to five pages selected for their well-organized caption–image structure.

Model comparison was conducted primarily among OpenAI’s GPT models, taking into account the objectives and efficiency of the PoC, and GPT-4o was judged to be the most suitable choice.<sup>51</sup> Additionally, setting the temperature value<sup>52</sup> to 0 yielded the highest response consistency, enabling its stable use for structuring tasks.

The time and cost required for caption matching were as follows. The Hwajisan report took approximately 504 seconds, with 309,425 input tokens (approximately \$0.77) and 46,516 output tokens (approximately \$0.47).

---

<sup>51</sup> The GPT-4o model was simply the most pragmatic choice for this paper’s PoC at the time of writing; it does not carry any particular technical or academic significance. Given that superior and more cost-effective models are highly likely to emerge in the future, readers are encouraged to focus not on the specific model itself but on the methodology of its application and the framework of the proof of concept.

<sup>52</sup> temperature is one of the key parameters of an LLM API, controlling the creativity or randomness of the output. Values closer to 0 produce more consistent and predictable responses, while values closer to 1 yield more diverse and creative outputs. For tasks requiring precision and consistency, such as data structuring, it is generally advisable to set a low value such as 0 or 0.1.

The Hangpaduri report took approximately 171 seconds, with 138,096 input tokens (approximately \$0.35) and 15,048 output tokens (approximately \$0.15). The Seokjangri report took approximately 390 seconds, with 302,061 input tokens (approximately \$0.76) and 30,549 output tokens (approximately \$0.31).<sup>53</sup>

Despite efforts to maximize accuracy, the caption matching results still exhibited limitations. In the Hwajisan report, 370 out of 1,156 images remained unmatched; in the Hangpaduri report, 27 out of 246; and in the Seokjangri report, 221 out of 706. The same task was also attempted using a programming approach based on regular expressions, which achieved excellent execution speed of under one second but still left unmatched images. The regex-based results likewise left 181, 26, and 217 images unmatched for the Hwajisan, Hangpaduri, and Seokjangri reports, respectively.

Consequently, neither approach is sufficiently accurate at the current level for direct use in practice. However, since the matched image and caption information was needed in subsequent modules during this verification process, it was used as-is despite some errors. While falling short of production-grade accuracy, the results secured sufficient coherence and consistency for the purpose of validating the overall pipeline flow. If, in the future, the regex logic is refined on the basis of a larger sample or accuracy is improved through iterative pattern learning and tuning, this approach could be expanded into a viable practical alternative.

### **(3) Automatic Metadata Extraction and Structuring Module**

Excavation reports are composed in the format of finished publications, and elements such as the cover, table of contents, preface, conclusion, and descriptions of the site's location and environment are often not primary targets for structuring. While these elements will ultimately need to be organized in a systematic manner, they were excluded from the analysis because their content is relatively straightforward and not directly connected to the core objectives of this proof of concept. This module limited its structuring to the descriptions of features and artifacts within the main body of the excavation report.

The data hierarchy was established in four tiers—site, trench, feature, and artifact—and an appropriate schema was designed manually, reflecting both the results of analyzing the three reports and archaeological domain knowledge. While it would be possible to delegate schema design entirely to the LLM, this would require considerable time, cost, and sophisticated tuning; accordingly, this proof of concept adopted an approach of clearly communicating a predefined schema and guiding the LLM to extract and organize information in accordance with it.

During the schema design process, the possibility of separating elements such as period, soil type, material, and feature/artifact types into distinct data types was also considered. However, this risked shifting the focus from structuring to the design of classification systems themselves, and since these elements can be adequately normalized through subsequent work, they were processed as natural language strings at this stage.

---

<sup>53</sup> The input/output processing speed of an LLM is influenced primarily by the type of model in use and the server conditions of the LLM provider, rather than by the performance of the user's device. Token costs can also vary considerably depending on the choice of model and the design of the pipeline and prompts.

The structuring task was performed on the basis of the output from the PDF processing module and the caption matching module constructed earlier, using GPT-4o as the LLM. While the results were satisfactory at the proof-of-concept level, the diversity of report formats presented more difficulties than anticipated. The Seokjangri report, for example, contained extensive descriptions of past excavations and surrounding Paleolithic sites in addition to the primary investigation, causing the LLM to confuse the main analytical targets with background information. This was addressed by explicitly restricting the scope of analysis with reference to the table of contents.

Moreover, the manner in which features and artifacts were described varied from report to report. Some reports described artifacts in connection with their associated features, while others separated features and artifacts and organized each in bulk. These differences in descriptive approach complicated prompt design, increased conditional logic, and heightened the contextual processing burden on the LLM, resulting in confirmed performance degradation. In cases where the text contained general overviews rather than individual descriptions of specific features or artifacts, the LLM sometimes treated these as a basis for structuring, producing results of diminished accuracy. To address this, prompts were configured by dividing the text into chunks—20-page segments for the general structuring of investigation sites and 2-page segments for detailed content on trenches, features, and artifacts—and through several dozen rounds of tuning, a certain level of consistency and accuracy was achieved.

For the Hwajisan report, the site overview and trench information were structured with relatively high accuracy and no significant omissions, yielding generally sound results. Feature and artifact information likewise exhibited stable structuring overall; although some instances of duplicate entities, omissions, and feature–artifact matching errors were identified, these were at a level that could be sufficiently improved with minor adjustments to the logic or prompts. Moreover, such errors were of a degree that could be relatively easily corrected by human reviewers in a subsequent verification process.

However, the linkage with images such as drawings and photographs was not complete. While no incorrect matches between entities and images were found, some items were missing, which could entail a considerable manual burden in a subsequent correction process. These issues are considered to have originated primarily from the incompleteness of the caption matching stage, and there is substantial room for natural improvement once the accuracy of that module is enhanced.

The time and cost required for metadata extraction and structuring in the final version were as follows. The Hwajisan report took approximately 34 minutes, with 5,492,541 input tokens (approximately \$13.73) and 111,831 output tokens (approximately \$1.12). The Hangpaduri report took approximately 5 minutes, with 239,651 input tokens (approximately \$0.60) and 21,185 output tokens (approximately \$0.21). The Seokjangri report took approximately 9 minutes, with 867,874 input tokens (approximately \$2.17) and 36,669 output tokens (approximately \$0.37).

#### **(4) Database Module**

This module is responsible for storing and managing the structured data produced by the preceding modules in a relational database (RDBMS). In an actual service deployment, it would be advisable to carefully select an RDBMS suited to archaeological data based on its characteristics and intended use; however, for this proof of concept, SQLite<sup>54</sup> was adopted in consideration of ease of implementation and speed. The module is organized around basic functions for storing refined data and querying necessary information, and its implementation was finalized at a level that performs simple input/output processing without complex logic.

**(5) Web Interface Module**

The web interface module was implemented with only the minimum functionality necessary for the proof of concept. Users can upload a PDF, and once the processed results are generated, the module provides list and tabbed table visualizations for browsing data by site (Figures 1–3). Originally, this module was intended solely for reviewing structuring results; however, in order to visually demonstrate how structured excavation report data could be utilized in an actual service, additional features were implemented. As these additions were somewhat different in character, they were separated out and organized as a distinct “AI Service Module.”



Figure 1. Web Interface Module: Site

<sup>54</sup> SQLite is a lightweight relational database management system (RDBMS) that requires no server and that operates by storing all data in a single file. Its ease of installation and operation makes it widely used in development and testing phases, and its ability to function without external dependencies is a significant advantage.

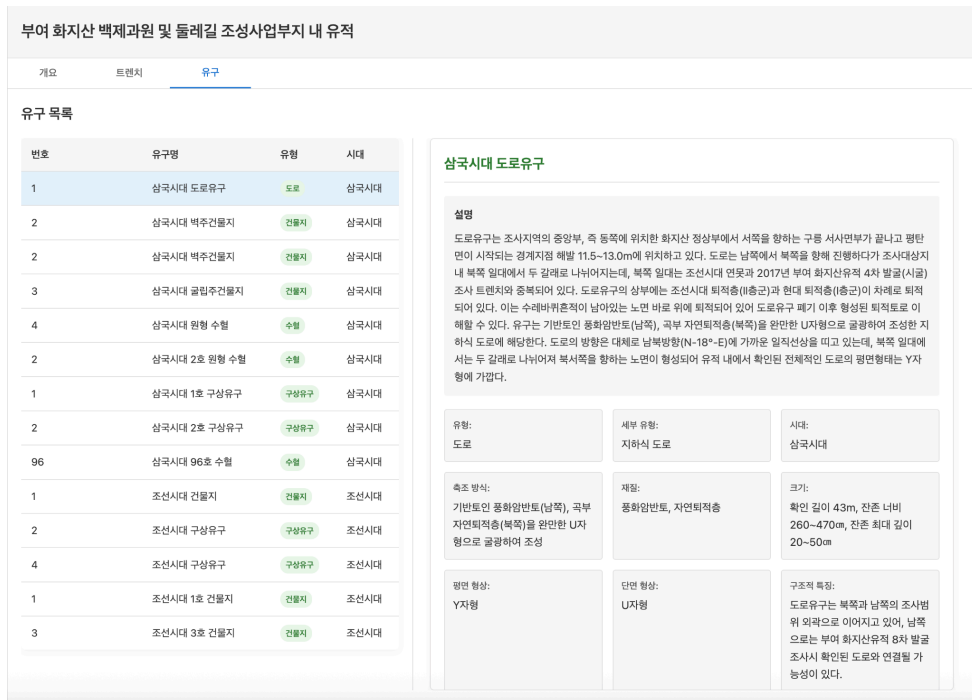


Figure 2. Web Interface Module: Features



Figure 3. Web Interface Module: Artifacts

(6) AI Service Module

The AI service module, built through the web interface, was designed to respond to users' natural language queries on the basis of the structured excavation report data. To enhance the accuracy and contextual

comprehension of the question-answering function, the contents of the *Dictionary of Korean Archaeology* were integrated via vector embedding<sup>55</sup> and RAG. This enables the LLM to generate responses grounded in more reliable background knowledge, while simultaneously providing users with a search environment in which information can be explored on a semantic basis rather than relying on keyword matching.

Through this, users can directly access organized information via queries such as “Tell me about the feature patterns found at the Hwajisan site” (Figure 4). In addition, a function was implemented to respond to requests for conceptual explanations so that general users unfamiliar with archaeology could also make use of the system. For example, even for questions such as “What is a *gullipju* (freestanding post structure)? The terminology is too difficult,” the LLM was configured to provide accessible explanations by referencing the *Dictionary of Korean Archaeology* and the report text (Figure 5).



Figure 4. AI Service Module: Feature Patterns at the Hwajisan Site

<sup>55</sup> Vector embedding is a technique that converts textual data—such as words, sentences, and documents—into high-dimensional numerical vectors, enabling the semantic similarity between natural language expressions to be represented numerically. Expressions sharing similar meanings or contexts are positioned closer together in vector space, thereby making semantic computation possible across diverse tasks such as search, classification, and clustering. In LLMs, vector embeddings are used to compare the semantic distance between a user query and reference documents, retrieving the most contextually similar content to enhance the accuracy of responses.

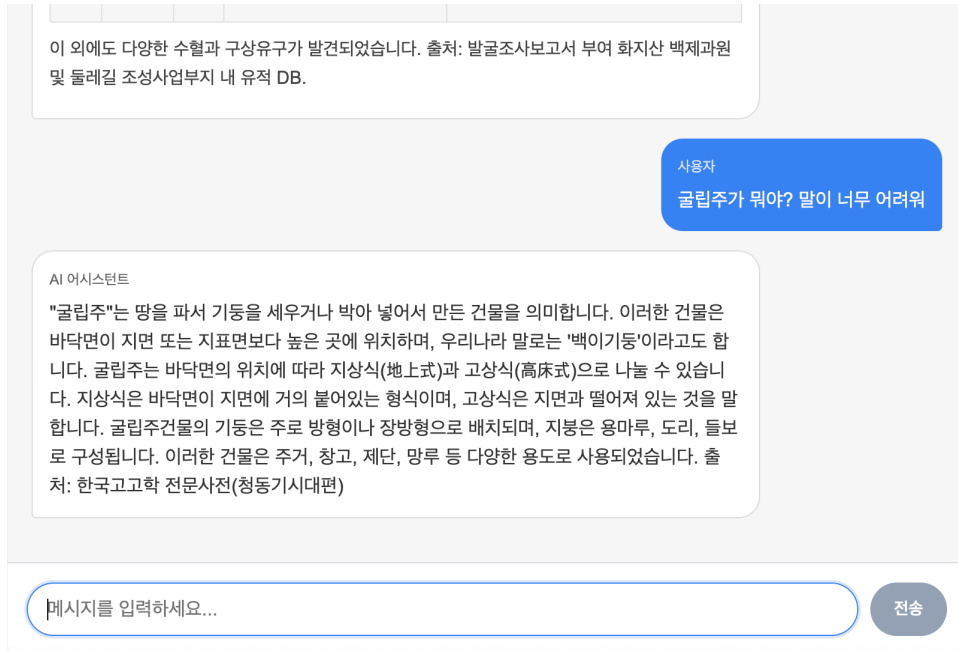


Figure 5. AI Service Module: Explanation of the Term “gullipju” (freestanding post structure)

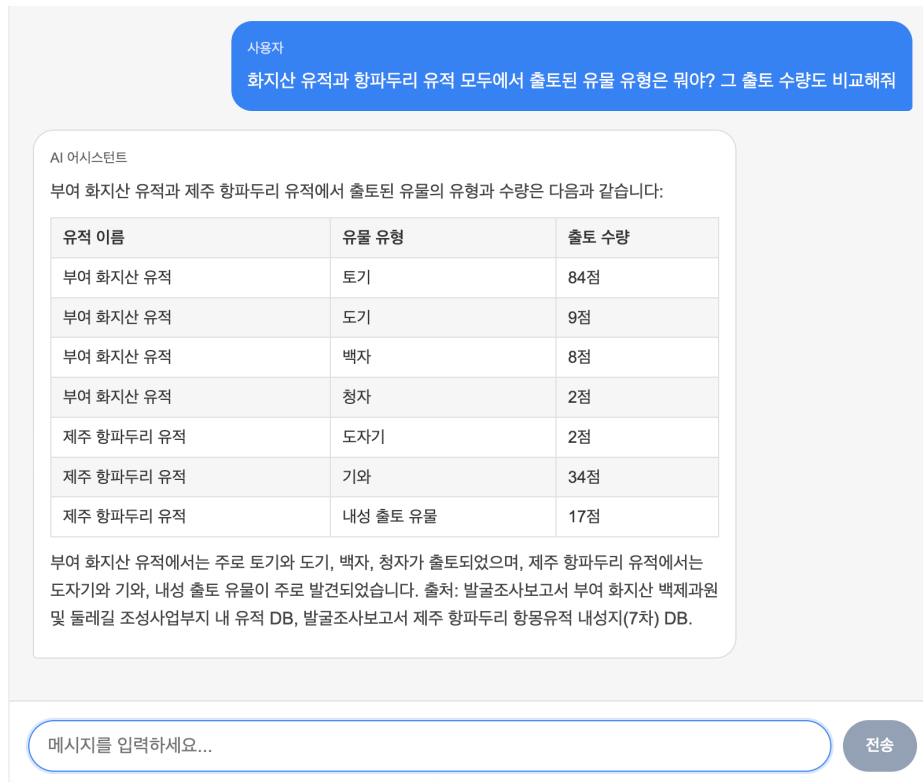
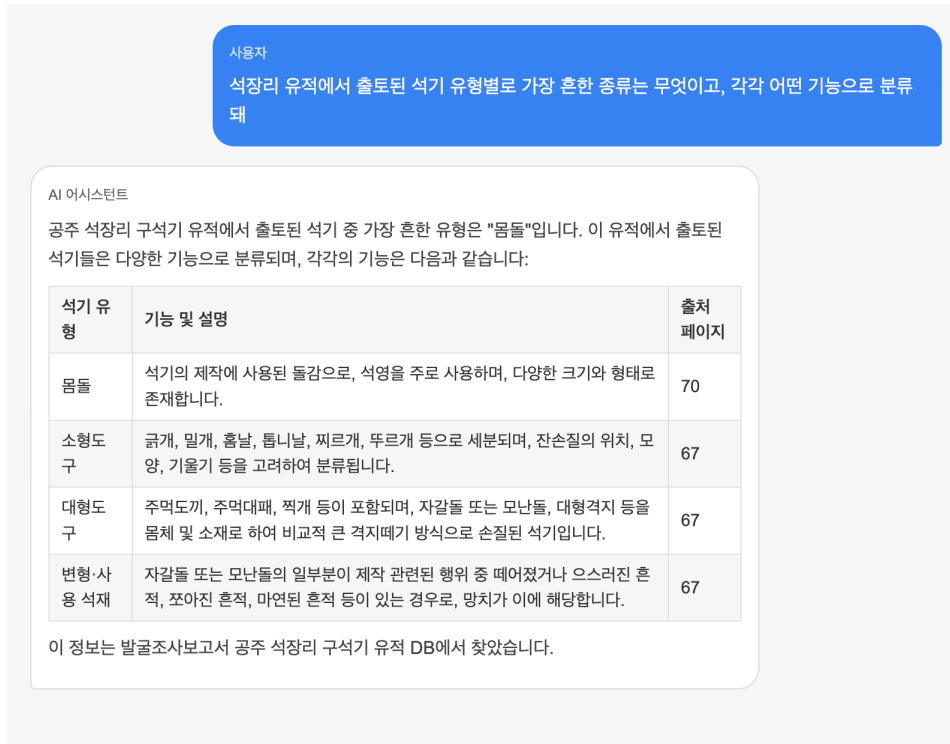


Figure 6. AI Service Module: Artifact Patterns from Hwajisan and Hangpaduri Sites



**Figure 7.** AI Service Module: Stone Tool Patterns at the Seokjangri Site

The module is also capable of handling comparative and analytical queries based on the structured data. For example, in response to the question “What artifact types were recovered from both the Hwajisan and Hangpaduri sites? Compare their quantities as well,” the system extracted common artifact categories and generated a response visualizing the quantitative differences in tabular form (Figure 6). Likewise, for the question “What are the most common types of stone tools recovered from the Seokjangri site, and how is each classified by function?” the system classified and organized artifact types and functions based on the report descriptions and returned a structured response (Figure 7).

Furthermore, even for information not explicitly recorded in the report database, the LLM was able to enhance the depth and quality of its responses by referencing the *Dictionary of Korean Archaeology* to supplement concepts or provide background information. For instance, in response to the query “Compare Silla-period pottery with Baekje-period pottery,” the system synthesized relevant entries from the dictionary and generated a response organizing the characteristics of pottery from both periods in tabular form by category (Figure 8). When asked about *seonmun tanal* (linear-pattern paddle-marking), the system even demonstrated the capacity to search for “seonmun” (linear pattern) and “tanal” (paddle-marking) separately and infer their combined meaning (Figure 9). In this way, the question-answering function combining structured report data with dictionary-based background knowledge demonstrated new possibilities for context-based search and analysis in the field of archaeology, extending well beyond simple information retrieval.

AI 채팅

사용자  
신라시대 토기와 백제시대 토기에 대해 비교해줘

AI 어시스턴트

신라시대와 백제시대의 토기는 각각 독특한 특징을 가지고 있습니다. 아래 표를 통해 두 시대의 토기를 비교해보겠습니다.

항목	신라시대 토기	백제시대 토기
시대 구분	고신라토기, 통일신라토기	한성기, 웅진기, 사비기
주요 특징	- 고신라토기는 4세기부터 통일 이전까지 사용됨 - 통일신라토기는 고신라 고유의 문화와 당나라 문물이 합쳐져 독특한 특징을 가짐 - 인화문토기의 성행, 유약 사용	- 한성기에는 연질-경질 기대가 공존하며 발형기대 발달 - 웅진기에는 연질계 기대 소멸, 발형기대 발달 - 사비기에는 통형기대의 호형부가 납작한 형태로 변화
주요 기종	- 굽다리접시, 목긴항아리 - 동물형토기, 토우	- 광구장경호, 발형기대 - 배형기대와 장경호의 조합
문양 및 장식	- 인화문, 연속마제형문, 국화문 등 - 문양이 화려하고 다양한 기종	- 투창, 고사리문 돌대 - 평면적 및 입체적 장식
기능 및 용도	- 일상 용기, 분묘 부장 토기, 장굴기	- 의례 용기, 제의 시 공헌기

출처:

- 신라시대 토기: 한국고고학사전(2001), 한국고고학 전문사전(고분유물편)
- 백제시대 토기: 한국고고학 전문사전(고분유물편)

진송

Figure 8. AI Service Module: Comparison of Silla and Baekje Pottery

사용자  
선문 타날이 뭐야?

AI 어시스턴트

"선문 타날"에 대한 직접적인 정의는 한국 고고학 사전에 없지만, "타날"과 "선문"에 대한 정보를 통해 유추할 수 있습니다.

- **타날**: 타날은 토기 제작 시 수분이 남아 있는 상태에서 그 표면을 두드려 외형을 다듬거나 기벽의 점토 조직을 치밀하게 하기 위해 사용하는 도구입니다. 타날 판을 사용하여 토기의 표면을 두드릴 때, 문양이 나타나게 됩니다. 이는 주로 토기의 외형을 다듬고 문양을 새기는 데 사용됩니다.
- **선문**: 선문은 토기나 기와의 표면에 선을 눌러 그은 문양을 의미합니다. 이는 주로 장식적인 목적으로 사용되며, 다양한 형태의 선문이 존재합니다.

따라서 "선문 타날"은 토기의 표면에 선문을 새기기 위해 타날 기법을 사용하는 것을 의미할 수 있습니다. 이는 토기의 장식성을 높이기 위한 기법으로, 토기의 외형을 다듬고 문양을 새기는 데 사용됩니다.

이 정보는 "한국고고학 전문사전(생산유적편)"에서 타날에 대한 설명을 참고하여 유추한 것입니다. 추가로 궁금한 점이 있으면 말씀해 주세요!

Figure 9. AI Service Module: Inferred Response to a Query on Seonmun Tanal (linear-pattern paddle-marking)

## 2) Evaluation and Directions for Improvement

Through this PoC, it was confirmed that the basic completeness of the structured output meets a certain threshold; however, various supplementary tasks remain for practical application. For the Hwajisan report, for example, the site overview and trench information were extracted with very good results, and feature and artifact information was also generally structured appropriately. Some instances of duplication between entities and errors in the linkage between features and artifacts were observed, and matching with images such as drawings and photographs was somewhat less consistent. This is primarily attributable to the incompleteness of the caption matching stage, and improving the accuracy of that stage would naturally enhance the quality of the overall structuring.

The descriptive approaches to features and artifacts varied across reports, necessitating the diversification of prompt design into conditional branching structures, which in turn increased the contextual processing burden on the LLM and in some cases led to performance degradation. A tendency was also observed in which the LLM attempted to use overview descriptions within the reports as a basis for structuring, requiring precise control. To address this, a page-level chunking strategy was introduced: general overviews of investigation sites were processed in 20-page segments, while detailed items such as features and artifacts were divided into 2-page segments. Through this tuning, a certain level of consistency and accuracy was secured.

Going forward, it will be necessary to refine the preprocessing pipeline itself: minimizing the complex conditional branching within prompts by analyzing the table-of-contents structure and contextual flow of reports to organize chunking around semantic units, and predefining branching strategies according to report type. Such an approach would reduce the contextual processing burden on the LLM while simultaneously improving both the precision of structuring and processing efficiency.

The costs associated with using commercial LLM APIs for the structuring task were not insignificant. Structuring alone incurred an API cost of approximately 20,000 KRW per report [Translation note: among the three test reports, the Hwajisan report was closest in volume to a typical Korean excavation report and was therefore used as the benchmark; the other two were excluded as clearly small-scale cases. The combined caption matching and metadata extraction costs for the Hwajisan report totaled approximately \$16.09 USD, converted at the prevailing exchange rate.], and if techniques such as self-reflection or RAG are employed in parallel, token consumption can increase by a factor of two to three. In a production environment processing several thousand reports annually, LLM API costs alone could reach hundreds of millions of KRW per year [Translation note: hundreds of millions of KRW corresponds to roughly several hundred thousand USD].

These cost concerns can be partially alleviated through the deployment of on-premises infrastructure<sup>56</sup> utilizing open-source LLMs,<sup>57</sup> the adoption of lightweight models, and prompt optimization. Furthermore, by minimizing the scope of high-cost LLM usage and employing strategies that involve domain-specific model tuning or the separation of processing stages, a balance between efficiency and accuracy can be struck.

For caption matching, both the LLM-based approach and the regex-based approach were attempted in parallel, but both exhibited somewhat high failure rates. Nonetheless, with continued iterative prompt tuning and pattern learning conducted in tandem, reaching a practical level of accuracy appears entirely feasible.

In this PoC, the structuring of feature and artifact information from report PDFs was achieved at a level exceeding a minimum threshold, and the source page from the original text was recorded alongside each structured entity, enhancing the practicality of verification and downstream use. When a traditional UI/UX-based information browsing function was implemented on the basis of this data, the result was a product enabling systematic exploration of information by site and by type. Furthermore, although not part of the original plan, an experimental extension into a natural language question-answering AI service module demonstrated that the function of responding to user queries on the basis of information stored in the structured database could be implemented in a stable manner. The system generated appropriate responses to a variety of queries, including comparisons of artifact types across sites, explanations of feature distribution patterns, and conceptual descriptions of specific structures.

In addition, by vectorizing the *Dictionary of Korean Archaeology* through embedding and integrating it via RAG, the system was able to provide contextually supplemented explanations even for concepts and terminology not present in the report data. This configuration holds considerable potential for extension beyond simple information responses to chart generation, similar artifact retrieval, and conceptual explanations, and confirmed the practical feasibility of implementing an advanced question-answering exploration environment incorporating semantic search. If, in the future, absolute coordinate data at the feature and artifact level, vector-format drawings, and 3D scan data become available, the realization of sophisticated functions such as map-based analysis, spatiotemporal visualization, and 3D modeling would also be entirely within reach.

Although the question-answering function in this PoC was implemented in a form that mimics the UI/UX of a typical AI chat service, I do not regard such a chat-based interface as necessarily suited to archaeological information services. The approach of having users directly input prompts can produce wide variation in response quality and consistency depending on the user's expressiveness and proficiency, limiting its capacity

---

<sup>56</sup> On-premises infrastructure refers to an information technology environment built and operated on physical servers or internal systems owned directly by the user, rather than utilizing cloud servers. Because data does not leave the organization, it offers advantages in security and control, making it the preferred choice for research institutions and public agencies handling sensitive information. However, it entails high initial deployment costs and requires technical expertise in hardware and software maintenance.

<sup>57</sup> An open-source LLM refers to a large language model distributed freely available for public use, based on publicly available code and model weights. Examples include Meta's LLaMA series, Mistral, Falcon, and BLOOM, all readily accessible through platforms such as Hugging Face. Compared to proprietary models, open-source LLMs offer advantages in flexibility and cost; however, depending on the model, there may be limitations in performance, fine-tuning scope, and the supporting ecosystem. They are attracting growing attention as alternatives to closed proprietary models, given that they can be deployed on on-premises servers or fine-tuned for specific domains as needed.

for stable and reliable information delivery. Therefore, if a dedicated UI/UX were developed that enables spatiotemporal visualization and purpose-driven exploration reflecting the distinctive characteristics of archaeology and cultural heritage, this could evolve into an unrivaled platform that achieves both domain expertise and accessibility.

Image-based PDF processing also remains a task to be resolved. For some actual reports, text extraction was impossible without OCR (optical character recognition). Although OCR was excluded at the PoC stage, image-based document processing technology incorporating OCR must be introduced for practical application.

A new approach to the design of classification systems also appears feasible. Whereas previously, structuring was carried out only after the establishment of a classification system, it is now entirely possible to derive a refined classification system in reverse from structured data and then batch-convert the data to conform to that system. This can be regarded as an advantage of the LLM-based approach in that it flexibly reconfigures the sequential relationship between informatization and the establishment of classification systems.

Finally, human verification is essential at the stage of practical application. A verification interface that presents the structured output and the original text in parallel is needed, and improvements must ultimately be directed toward raising the reliability of the structuring module itself so as to reduce the verification burden. By striking a balance among accuracy, speed, and cost, and by designing a structure amenable to incremental refinement and scalable expansion, archaeological informatization can take one step closer to viable field application.

## V. Conclusion

This paper has focused on systematically organizing the discourse on informatization that has emerged sporadically in the field of archaeology, and on establishing a conceptual foundation for informatization and the utilization of artificial intelligence. As the first comprehensive discussion of informatization attempted within Korean archaeology, it may serve as a reference point for future related research.

At present, archaeological data analysis largely remains at the level of individual researchers conducting ad hoc experiments on limited datasets. However, if the informatization framework proposed in this paper is realized, data-driven analysis could become established as a more routine research tool. While this paper has focused on examining a conceptual model and its validity, the next step will be to advance toward the development of an operationally viable system on this basis. I am currently planning a web-based platform built on this model, and intend to release a prototype service in the second half of 2025 that archaeologists can use directly. The service will provide functions such as automatic report analysis, structuring of feature and artifact data, and retrieval of similar cases, and will be continuously refined through user feedback.

Just as New Archaeology enhanced the analytical precision of the discipline through the introduction of scientific methodology, and contextual archaeology foregrounded the importance of cultural interpretation, so “Informational Archaeology,” grounded in the process of archaeological informatization proposed in this paper, may constitute a new paradigm for archaeology in the digital age. This is an approach that transcends the mere adoption of technology, with the potential to transform the entirety of knowledge production, access, and interpretation, and it may make a substantive contribution to future archaeological research. To be sure, it is premature to assert the emergence of a new paradigm on the basis of a single proof-of-concept study alone. Nonetheless, insofar as the automatic structuring of data and intelligent exploration presented in this paper enable an approach qualitatively distinct from existing research methods, I wish to propose the possibility of a new paradigm called “Informational Archaeology.”

A digital transformation in the field of archaeology, accompanied by informatization, will soon become a reality. In particular, enhancing the quality of structured information requires not only the acquisition of primary digital resources such as drawings and 3D scans but also the establishment of institutional mechanisms for their systematic deposit and public release. At present, most archaeological data is confined to reports in PDF<sup>58</sup> format, and even these vary widely in form and quality, limiting their reproducibility and utility. To enhance the practical effectiveness of archaeological informatization going forward, policy discussions must proceed in parallel—addressing the integrated collection and release of primary data, the formulation of standardized submission rules, and the coordination of inter-institutional collaboration frameworks. In this regard, it has

---

<sup>58</sup> Some excavation reports are still produced as image-based PDFs, posing a significant obstacle to informatization. Older reports are predominantly raster-format scans of printed volumes, but recent cases have also been identified in which the entire text and body are converted to outlines and produced in vector format, ostensibly to prevent font substitution or ensure print registration accuracy. Such machine-unreadable formats necessitate OCR (Optical Character Recognition) in the informatization workflow, resulting in quality degradation and increased costs. In the long term, it would be desirable to submit machine-readable, structured data separately from print-oriented reports; however, the current report system is likely to persist for the foreseeable future. At a minimum, future report production should incorporate basic design considerations for machine readability and digital processing—such as refraining from converting text to outlines.

recently been argued that the roles of metropolitan and municipal governments must be clearly defined, and that the procedures spanning information generation, transmission, and preservation should be codified at the ordinance level (Gang and Joo 2024). On such an institutional foundation, technology can take root more firmly, and the sustainability of informatization can be assured.

The theoretical and technical foundations proposed in this study are not confined to archaeology alone; they can ultimately be extended to the entire domain of cultural heritage. Although each field possesses its own distinct characteristics, the principles of systematic structuring and intelligent utilization of information constitute a universal foundation applicable across all of them. Such an integrative approach can serve as a basis for deepening our understanding of the interconnections and contexts among different areas of cultural heritage. It is hoped that this work may mark the starting point of a new current in future discussions on the digital transformation of cultural heritage—one that places informatization at its central axis.

#### ※ Digital Appendix (<https://poc.heripo.com>)

The PoC process discussed in Chapter IV was introduced through selected screenshot images; however, the constraints of a print medium alone cannot fully convey the overall workflow and operational dynamics. A separate digital appendix has therefore been prepared. On this webpage, readers can view video recordings of the PoC implementation process as it stood at the time of writing and submission. The appendix is intended to be continuously updated in an increasingly refined form as the underlying technology advances. By the time of publication, it is possible that substantially more polished results—addressing many of the technical and practical limitations revealed in the PoC—will already be available.

Readers wishing to examine the latest progress and practical implementations of this research are therefore strongly encouraged to consult the digital appendix. Depending on the time of access, interactive features may also be available in addition to video content. [Translation note: As of April 2026, when this translation was prepared, the digital appendix has been updated well beyond the PoC stage described in this paper. The system's open-source successor, heripo engine (<https://github.com/heripo-lab/heripo-engine>), is actively under development and is listed on Open Archaeo (<https://open-archaeo.info/post/heripo-engine>), an international directory of open-source archaeological software.]

### **Acknowledgments for the Translated Edition**

I am deeply grateful to the National Research Institute of Cultural Heritage (국립문화유산연구원) and its staff for graciously granting permission for this derivative work, making it possible for this translated edition to see the light of day. I also wish to thank CHO Hayoung and KIM Gaeun of heripo lab, who share the same dream and are devoted to research and development alongside me.

I extend my special thanks to The Korean Archaeological Society under its 31st presidency for willingly embracing the challenge of a single individual and joining in the journey toward the digital transformation of Korean archaeology.

My deepest respect goes to Dr. CHOI In-hwa—the “mother of Korean digital archaeology” and my first excavation mentor—for recognizing the value of my work despite my lack of formal credentials, and for bringing it to the wider world.

Finally, I dedicate this paper to my beloved wife, KIM Bobae, and my son, KIM Minjun, who quietly yet wholeheartedly support my selfish journey of self-fulfillment.

## References

**Note:** Official English bibliographic information was prioritized. In the absence of an official English title, reliable sources were consulted; for instance, authors' romanized names were verified against their other peer-reviewed publications. For terms related to the International Symposium: Digital Transformation of Archaeological Information, the official English press releases from the Korea Heritage Service (KHS) were utilized. All English titles for works without official translations were provided by the author.

### Journal Articles and Conference Proceedings

- AHN Hyoung-Ki (안형기). 2017. 'The Research on Convergence in Information and Communication Technology (ICT) applied to Archaeology' [考古學과 ICT 融合活用 研究]. Ph.D. dissertation, *Korea University* [고려대학교], <http://www.dcollection.net/handler/korea/000000077125>.
- Brown, T. B. et al. 2020. 'Language models are few-shot learners'. *Advances in neural information processing systems*, 33, <https://doi.org/10.48550/arXiv.2005.14165>.
- Caren, N. et al. 2024. 'Extracting Protest Events from Newspaper Articles with ChatGPT'. <https://doi.org/10.31235/osf.io/dvht7>.
- CHA Soon Chul (차순철). 2016. 'Archeological Materials Database Development, Current Status and Tasks' [고고학 자료 데이터베이스화의 현황과 과제]. *The Journal of Korean Ancient History* [한국고대사연구], 84, 75–119, <https://uci.kci.go.kr/resolver/G704-000848.2016..84.004>.
- CHOI In-Hwa (최인화). 2022. 'A Study on the vitalization of Digital Archaeology based on European case studies' [디지털고고학 활성화 방안 연구 - 유럽 사례를 중심으로 -]. Ph.D. dissertation, *Pusan National University* [부산대학교], <https://dcollection.pusan.ac.kr/common/orgView/000000153918>.
- CHOI In-Hwa (최인화). 2023. 'Digital Archaeology Case Studies Overseas : Current Status and Challenges' [디지털고고학 국외 사례 현황과 과제]. *The Journal of Korean Field Archaeology* [야외고고학], 48, 113–154, <https://doi.org/10.35347/jkfa.2023..48.113>.
- CHOI In-Hwa (최인화). 2024. 'Research Trends and Future Tasks in the Field of Korean Digital Archaeology' [디지털 기술을 활용한 국내 고고학 분야의 연구 동향과 향후 과제]. *Digital+Humanities*, 30, 71–89.
- GANG Bobae (강보배) and JOO Sanghun (주상훈). 2024. 'A Study on How to Improve Local Government Ordinances to Promote Korean Heritage Information Management: Focusing on the relation aspect of Korean Heritage laws and ordinances' [국가유산 정보 관리 활성화를 위한 광역자치단체 조례의 개선 방향 연구: 국가유산 관련 법과 조례의 관계적 성격을 중심으로]. *Korean Journal of Heritage: History and Science* [헤리티지: 역사와 과학], 57(3), 140–160, <https://doi.org/10.22755/kjchs.2024.57.3.140>.
- JUNG Ik-Jae (정익재). 2020. 'A Study on the Improvement of Excavation and Research Process - With a Focus on Building a Silla Ancient Tombs Database -' [문화재 발굴 조사·연구 과정의 개선 방안 연구- 신라 고분 데이터베이스 구축을 중심으로 -]. *Korean Journal of Cultural Heritage Studies* [文化財], 53(3), 4–23, <https://doi.org/10.22755/kjchs.2020.53.3.4>.
- KANG Dongseok (강동석). 2022. 'Transformation of the Dictionary of Korean Archaeology: Establishment of a GIS-based Knowledge Platform' [한국고고학사전의 대전환, GIS 기반 지식정보 플랫폼 구축]. *International Symposium Digital Transformation of Archaeological Information; [국제학술심포지엄 고고학술정보 디지털 대전환]*, National Research Institute of Cultural Heritage, 171–190.
- KANG Dongseok (강동석). 2024. 'Proposal of WebGIS-based Korean Archaeological Dictionary Information Service Model' [WebGIS 기반 한국고고학사전 정보서비스 모델의 제안]. *Korean Journal of Heritage: History and Science* [헤리티지: 역사와 과학], 57(1), 6–19, <https://doi.org/10.22755/kjchs.2024.57.1.6>.
- KIM Bumcheol (김범철). 2023. 'Strategies and Challenges in Digitizing Archaeological Data' [고고 디지털 아카이브 구축의 과제와 전략]. *Korean Journal of Cultural Heritage Studies* [文化財], 56(1), 6–18, <https://doi.org/10.22755/kjchs.2023.56.1.6>.
- KIM Hongyeon (김홍연). 2013. 'A Review of Digital Informatization Strategies for Buried Cultural Heritage' [매장문화재 디지털 정보화 방안 검토]. *Collection of 2013 Graduation Dissertations, Department of*

- Archaeology, Korea National University of Heritage [2013학년도 한국전통문화대학교 문화유적학과 졸업논문발표집]*, 284–298. (Translated by the author)
- KIM Hongyeon (김홍연). 2015. ‘A Proposal for Utilizing Big Data for Efficient Cultural Heritage Management’ [효율적 문화재관리를 위한 빅데이터 활용방안]. *Proceedings of the 2015 Winter Conference of the Seoul Association for Public Administration [2015년 서울행정학회 동계학술대회 발표논문집]*, Seoul Association for Public Administration, 317–332. (Translated by the author)
- KIM Hoon-Hui (김훈희). 2023. ‘Digital Recording of Archaeological Data: Past, Present, and Future of Artifact Drawings’ [고고자료의 디지털 기록: 유물도면의 과거·현재·미래]. *Proceedings of the 47th National Congress of the Korean Archaeological Society [제47회 한국고고학전국대회 발표자료집]*, The Korean Archaeological Society, 659–670. (Translated by the author)
- KIM Hyeon (김현) and KANG Haewon (강혜원). 2022. ‘Strategies and Implementations for Archiving Archaeological Information’ [고고학술정보 데이터 아카이빙 전략과 구현 방안: 한국고고학사전의 디지털 전환을 위한 시맨틱 데이터 아카이브 제언]. *International Symposium Digital Transformation of Archaeological Information; [국제학술심포지엄 고고학술정보 디지털 대전환]*, National Research Institute of Cultural Heritage, 125–148.
- KIM Ji-In (김지인). 2000. ‘A Study of Relational Database & Object Orient Database for Artifact Information Management’ [유물 정보 관리 데이터베이스의 관계형 모델과 개체 지향형 모델 비교 연구 : 모델 KJI-ADB1과 모델 KJI-ADB2 제시]. Master's thesis, *Seoul National University [서울대학교]*, <https://hdl.handle.net/10371/32464>.
- LEE Chang-hee (이창희). 2023. ‘Musing on Archaeological Digital Record -indoor-’ [고고학적 디지털 기록에 대한 단상(斷想)-indoor-]. *The Journal of Korean Field Archaeology [야외고고학]*, 48, 155–173, <https://doi.org/10.35347/jkfa.2023.48.155>.
- LEE Chorong (이초룡). 2022. ‘The Past and Present of Establishing Archaeological Information System’ [고고학술정보 구축의 어제와 오늘: 한국고고학사전 구축 사업을 중심으로]. *International Symposium Digital Transformation of Archaeological Information; [국제학술심포지엄 고고학술정보 디지털 대전환]*, National Research Institute of Cultural Heritage, 7–30.
- LEE Chorong (이초룡). 2023. ‘Approaches to Creating a Digital Encyclopedia of Korean Archaeology’ [한국고고학 디지털 사전구축 방안 연구]. *Korean Journal of Cultural Heritage Studies [文化財]*, 56(2), 28–45, <https://doi.org/10.22755/kjchs.2022.56.1.28>.
- LEE Minseok (이민석). 2024. ‘Research on the characteristics and utilization of buried cultural heritage - Focusing on Jeonbuk region -’ [매장유산의 특징과 활용방안 연구 - 전북지역을 중심으로 -]. *Korean Journal of Heritage: History and Science [헤리티지: 역사와 과학]*, 57(4), 240–267, <https://doi.org/10.22755/kjchs.2024.57.4.240>.
- Ntinopoulos, V. et al. 2025. ‘Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation’. *BMJ Health & Care Informatics*, 32(1), <https://doi.org/10.1136/bmjhci-2024-101139>.
- Ouyang, L. et al. 2022. ‘Training language models to follow instructions with human feedback’. <https://doi.org/10.48550/arXiv.2203.02155>.
- Raza M. Z. et al. 2025. ‘LLM-TA: An LLM-Enhanced Thematic Analysis Pipeline for Transcripts from Parents of Children with Congenital Heart Disease’. <https://doi.org/10.48550/arXiv.2502.01620>.
- RHEE Hea Lim (이혜림). 2018. ‘Developing the Korean National Archaeological Data Digital Archive: An Exploratory Study’ [국가 고고학 데이터 디지털 아카이브 개발을 위한 연구]. *Journal of Korean Society of Archives and Records Management [한국기록관리학회지]*, 18(2), 1–28, <https://doi.org/10.14404/JKSARM.2018.18.2.001>.
- SO Jaeyun (소재윤) and KIM Juho (김주호). 2025. ‘Current Status and Policy Recommendations for Managing Excavation Records in the Mahan Historical and Cultural Region - Focusing on the Mahan Historical and Cultural region -’ [발굴기록물 보존관리 현황과 정책 제언 - 마한역사문화권 현황을 중심으로 -]. *Korean Journal of Heritage: History and Science [헤리티지: 역사와 과학]*, 58(1), 22–43, <https://doi.org/10.22755/kjchs.2025.58.1.22>.
- Vaswani, A. et al. 2017. ‘Attention is all you need’. <https://doi.org/10.48550/arXiv.1706.03762>.

- YIM Jeong Bin (임정빈). 2011. ‘A Study on a Plan of Records Management for Excavation Research Records - On the Basis of Buried Cultural Properties -’ [발굴조사 기록의 기록관리 방안 연구 - 매장문화재를 중심으로 -]. Master's thesis, Myongji University [명지대학교], <https://mju.dcollection.net/common/orgView/200000923885>.
- Yuichi Takada and Soyeon Ho. 2022. ‘Discovering the Archaeological Sites throughout Japan’ [일본 전국의 유적을 검색하다], *International Symposium Digital Transformation of Archaeological Information; [국제학술심포지엄 고고학술정보 디지털 대전환]*. National Research Institute of Cultural Heritage, 31–54.
- Zhao, W. X. et al. 2023. ‘A survey of large language models’. <https://doi.org/10.48550/arXiv.2303.18223>.
- Zininzin Academic Team (진인진 학술팀). 2005. ‘Issues in the Informatization of Buried Cultural Heritage’ [매장문화재 정보화의 제 문제]. *Journal of Honam Institute of Cultural Heritage [호남문화재연구원 연구논문집]*, 5, 173–188. (Translated by the author)

## Reports and Monographs

- Baekje History and Culture Institute (백제문화유산연구원). 2025. *The Baekje Garden and Trail Site in Hwajisan Mountain, Buyeo* [부여 화지산 백제과원 및 둘레길 조성사업부지내 유적].
- Cultural Heritage Administration (문화재청). 2012. *Five-Year Basic Plan for the Preservation, Management, and Utilization of Cultural Heritage* [문화재 보존·관리·활용 5개년 기본계획]. (Translated by the author)
- Cultural Heritage Administration (문화재청). 2013. *Plan for Cultural Heritage 3.0* [문화유산 3.0 추진계획]. (Translated by the author)
- Cultural Heritage Administration (문화재청). 2021. *Digital Transformation Plan for Cultural Heritage (Press Release: Announcing ‘Digital Transformation of Cultural Heritage 2030’)* [문화재 디지털 대전환 계획. 문화재청, 문화재청 보도자료 ‘문화재 디지털 대전환 2030’ 발표-올해부터 2030년까지 문화재 보존·관리·활용 전 분야에 디지털 방식 도입- 불임1]. (Translated by the author)
- Digital Antiquity. 2022. *The Digital Archaeological Record (tDAR)*.
- Gyeongju Institute of Cultural Heritage (경주문화유산연구원). 2025. *Seokjangri Paleolithic Site(14th), Gongju* [공주 석장리 구석기 유적(14차)].
- Il-young Cultural Heritage Research Institute (일영문화유산연구원). 2025. *Jeju Hangpaduri Hangmong Antimeral Resistance Site (7th)* [제주 향파두리 항몽유적 내성지(7차)].
- National Research Institute of Cultural Heritage (국립문화유산연구원). 2024. *2024 National Research Institute of Cultural Heritage Brochure(Korean)* [2024 국립문화유산연구원 홍보책자(국문)].
- National Computerization Agency (한국전산원). 2000. *A Study on Evaluating Information Level in the Korean Traditional Culture Area* [문화 정보화수준 평가 연구 - 전통문화유산분야를 중심으로].

## Websites

- “ChatGPT”. Wikipedia. Accessed April 30, 2025, <https://en.wikipedia.org/wiki/ChatGPT>.
- “Current Status of Buried Heritage Excavations (e-Nara Index, 2020–2024)” [매장유산 발굴조사 현황(e-나라지표, 2020~2024년)]. Korea Heritage Service. Accessed April 30, 2025. [https://cha.go.kr/html/HtmlPage.do?pg=/seek/search/statistics-02\\_05.jsp&mn=NS\\_03\\_07\\_02](https://cha.go.kr/html/HtmlPage.do?pg=/seek/search/statistics-02_05.jsp&mn=NS_03_07_02). (Translated by the author)
- “NRICH Archive” (국가유산 지식이음). National Research Institute of Cultural Heritage. Accessed May 4 2025, <https://portal.nrich.go.kr>.
- “The Crossroads Developers Face: Learn Vibe Coding, or Retire” (개발자가 맞닥뜨린 갈림길 ‘바이브 코딩’을 배우거나, 은퇴하거나). ITWorld Korea. Accessed May 4, 2025, <https://www.itworld.co.kr/article/3967678>. (Translated by the author)