Research Square

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

Embodied-State-Driven Agentic GraphRAG for Contextual Interpretation of Ancient Chinese Alchemical Heritage

Yumiao Chen

cym@ecust.edu.cn

East China University of Science and Technology

Tanghua Xie

East China University of Science and Technology

Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11615622/v1

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

Additional Declarations: No competing interests reported.

Embodied-State-Driven Agentic GraphRAG for Contextual Interpretation of Ancient Chinese Alchemical Heritage

Yumiao Chen[1*] and Tanghua Xie[1]

[1*] School of Art, Design and Media, East China University of Science and Technology, Shanghai, 200000, China.

*Corresponding author(s). E-mail(s): cym@ecust.edu.cn;

Abstract

Ancient Chinese alchemy is an important part of scientific and technological heritage, preserving historical knowledge of materials, apparatus and technical practices. However, fragmented sources and largely scripted digital interpretations make it difficult to connect these practices with reliable evidence and modern scientific understanding during interaction. We developed an extended reality (XR) system integrating a provenance-aware knowledge graph with Agentic GraphRAG. Its Embodied State as Query mechanism uses process stages, objects, actions and material states to retrieve historical passages and contextual scientific explanations. Agentic GraphRAG achieved the highest scores across four offline retrieval and interpretation measures. In a study of 66 participants, adaptive XR yielded greater conceptual and procedural knowledge gains than web-based learning. Compared with scripted XR, it received higher ratings for explanation credibility and source clarity and elicited more mechanism-oriented questions. These findings support state-conditioned, evidence-grounded interpretation of historical technical practices during interaction.

1. Introduction

Ancient Chinese alchemy preserves a substantial record of materials, apparatus and technical practices within China's scientific and technological heritage [1]. Its texts describe mineral preparation, heating, separation and observation, often using terminology and explanations specific to their historical settings. Historians of alchemy have examined these practices as forms of knowledge production shaped by experiment, material manipulation and contemporary ideas about matter. Moran places alchemy within the history of science [2], while Principe [3] and Newman [4] examine its historical interpretation and changing historiography. Chang further discusses the relationship between material inquiry, life and vitalism in early modern chymistry [5]. Read together, these studies suggest that the significance of alchemical heritage lies partly in the procedures and observations that connected materials, instruments and historical explanations.

Explaining historical alchemical practices to present-day audiences requires attention to both the original accounts and the scientific interpretation of material phenomena. Research on chemical symbolism [6], historical mercury-related experiments [7] and the composition and alteration of mercury compounds [8] provides different perspectives on these practices. Historical texts establish what was recorded and how an operation was understood at the time; modern chemistry helps explain the material behavior involved. An

interactive interpretation system must make these sources of knowledge accessible while keeping their different evidential roles clear.

Digital heritage research has examined virtual museums [9], three-dimensional reconstruction [10], embodied augmentation [11], virtual narrators [12] and immersive visitor experiences [13]. Reviews of museum learning [14] and cultural heritage applications [15] have also assessed the educational uses of VR, AR and other immersive technologies. These studies show how digital environments allow visitors to inspect and interact with reconstructed objects. For historical scientific apparatus, interpretation also depends on explaining the purpose of an operation, the material changes associated with it and the documentary evidence supporting that account.

Studies of museum gamification [16], gesture-based cultural activities [17] and embodied information spaces [18] have examined how physical participation shapes engagement with cultural content. Learning-oriented XR research further connects sequences of user actions with task-specific feedback [19]. In the present study, these findings motivate a more specific use of interaction data: the operation being performed becomes part of the query used to retrieve an explanation.

Context-aware museum guides such as iMuseum [20], UbiCicero [21] and SMARTMUSEUM [22] use location and activity information to select content for visitors. Related work has linked instructional support to scientific tasks [23], recognized user actions in augmented reality [24] and adapted feedback during hands-on virtual learning [25]. These approaches establish the value of context-sensitive delivery. The question addressed here is how an observed operation can also determine which historical passages, technical relationships and scientific explanations are retrieved.

Knowledge graphs have been used to organize heritage collections [26], describe traditional crafts [27], and answer questions about Nanjing Yunjin [28] and historical clothing [29], as well as support multimodal access to painting and calligraphy [30]. More recent applications link graph relationships to ceramic interpretation [31], multimodal museum retrieval [32], source-aware question answering [33] and temporal reasoning [34]. These systems provide ways to represent and retrieve relationships among cultural objects and their associated knowledge. The present study examines a different point of access: selecting evidence according to the process stage and object with which a visitor is currently interacting.

Graph-based knowledge representation becomes especially useful when historical expressions and modern scientific categories cannot be treated as interchangeable. To avoid presenting a reconstructed apparatus or contemporary chemical explanation as a direct historical fact, the interpretation process must retain source attribution, represent uncertainty and distinguish documented from inferred features. This provides the heritage-specific motivation for combining semantic asset binding with traceable retrieval and source checking.

Retrieval-augmented language models use external documents to inform generated responses, as established by REALM [35], dense passage retrieval [36] and RAG [37]. GraphRAG adds access to relationships among retrieved entities and has been studied for query-focused summarization [38] and other graph-based retrieval tasks [39]. Agentic RAG introduces task-dependent retrieval planning and evidence checking [40]. In these approaches, retrieval is generally initiated by a text query. We examine how an XR interaction state can supply additional query information, allowing the system to select evidence associated with the current object,

process stage and material condition.

This study examines how the state of an ongoing XR activity can guide the retrieval of historical evidence and scientific explanations. A learner handling a reconstructed vessel may need information about its historical function, its place in a procedure or the material phenomenon currently being simulated. These requests depend partly on what the learner is doing when the question arises. We therefore investigate the use of interaction state as retrieval context and compare adaptive feedback with scripted XR and Web-based learning.

We call this mechanism Embodied State as Query. It encodes the current process stage, active object, recent action, material state and user question as a structured query. Agentic GraphRAG uses these fields to retrieve relevant passages and graph relationships, checks the supporting evidence and generates an explanation for the current task. The XR interface presents the explanation with its historical sources and corresponding modern scientific account.

We evaluated the system through a virtual reconstruction of a historical alchemical process involving cinnabar. Participants assembled digital apparatus, adjusted simulated parameters and observed material-state changes in XR. The interpretation content combined historical documentary records with separately attributed modern chemical explanations. Evaluation included an offline comparison of three retrieval methods and a 66-participant study measuring knowledge gains, workload, user experience and information-seeking behavior.

The contributions are threefold:

(1) A provenance-aware knowledge organization approach that relates historical texts, apparatus evidence, procedural steps, material states and modern scientific concepts while retaining source distinctions and uncertainty.

(2) An Embodied State as Query mechanism that uses current XR actions and process states to select graph-linked documentary evidence and contextual explanations through Agentic GraphRAG.

(3) A paired evaluation of offline interpretation quality and learner outcomes, including a comparison between dynamic and scripted feedback within the same XR environment. This design distinguishes evidence-grounded interaction quality from learning gains that may also arise from immersion itself.

2. Methods

2.1 Overall Framework

The system comprises three layers: semantic modeling, intelligent generation and interactive presentation (Fig. 1). Semantic modeling organizes historical texts, apparatus records and scientific explanations in a knowledge graph. The intelligent generation layer links reconstructed objects to graph entities and retrieves evidence according to the current interaction state. The presentation layer records users' actions in XR and displays the resulting explanations. Together, these components allow the system to retrieve information about a specific apparatus or process step when that object or step becomes the focus of interaction.

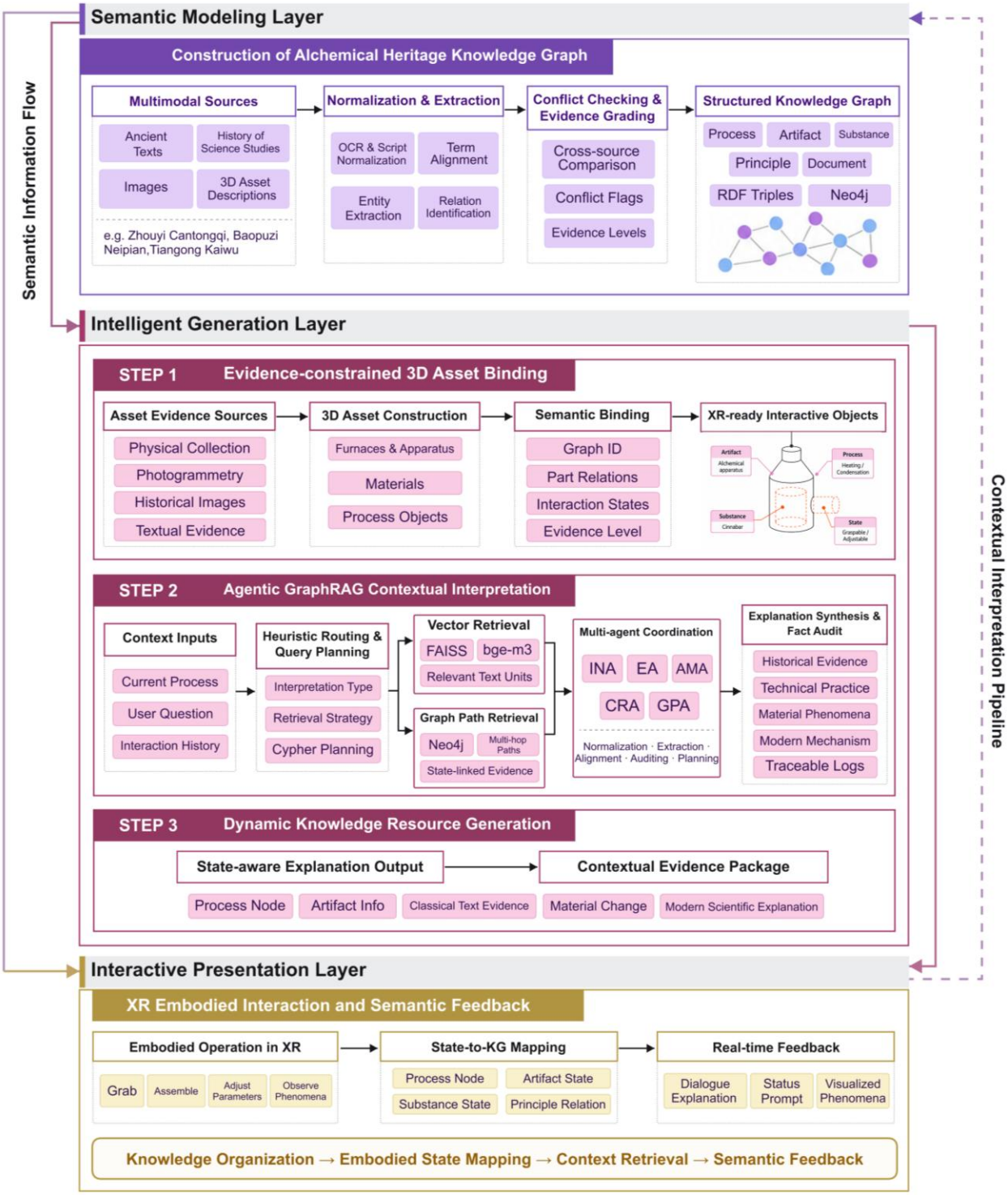


Fig. 1 Overall framework for embodied heritage interpretation

The Semantic Modeling Layer constructs a domain knowledge graph of processes, apparatus structures, materials, documentary sources and scientific concepts associated with ancient Chinese alchemy. Inputs include the Zhouyi Cantongqi, Baopuzi Neipian, histories of science, visual records and 3D asset descriptions. Terminology normalization, entity extraction, relation identification, conflict checking and evidence grading produce entity–relationship statements stored in a Neo4j property graph. Source metadata and provenance

links support path-based retrieval, terminology mapping and evidence-grounded feedback.

The Intelligent Generation Layer converts structured heritage knowledge into contextual resources that can be presented and accessed during interaction. These include 3D apparatus assets and a dynamic interpretation mechanism based on Agentic GraphRAG. Descriptions of apparatus, materials, steps, historical evidence and modern mechanisms inform the construction of interactive spatial objects. State recognition, query planning, vector retrieval, graph-path retrieval, explanation synthesis and factual auditing connect historical technical knowledge to modern scientific explanations within the current operational context.

The Interactive Presentation Layer embeds the semantic network and digital assets in an XR interface. Users grasp, assemble and adjust virtual apparatus and materials through spatial gestures. These actions are mapped in real time to graph entities, process nodes, material states and mechanistic relationships, triggering retrieval and reasoning. Dynamic dialogue, state prompts and visualized phenomena provide feedback, connecting embodied action with access to historical evidence and explanations of modern scientific mechanisms.

In operation, each recognized XR event updates the semantic state used for retrieval. The selected passages and graph relationships are passed to the generation module, and the resulting explanation is displayed alongside the current task. Subsequent user actions update the query context for the next request.

2.2 Semantic modeling layer

The source material includes historical texts, scientific histories, apparatus images and three-dimensional records. Differences in terminology, character forms and document editions require normalization before the records can be linked. The semantic modeling process therefore consists of three stages: source normalization, multi-agent extraction and checking, and knowledge graph construction. Source identifiers and edition information are retained throughout processing.

2.2.1 Multimodal historical data ingestion and normalization

The multimodal corpus comprises historical and scholarly texts, two-dimensional visual records and three-dimensional digital assets associated with ancient Chinese alchemical practices. Textual sources document technical procedures, apparatus functions and material phenomena; visual records provide evidence of apparatus forms, component arrangements and process configurations; and 3D assets represent the spatial structures and interactive properties of reconstructed objects. Together, these materials provide complementary documentary, visual and spatial information for semantic modeling.

The historical corpus draws on representative texts and studies of Chinese alchemy, including Science and Civilisation in China [41], Chinese Alchemy: Preliminary Studies [42], Great Clarity [43] and a history of ancient Chinese chemical technology [44], alongside the Zhouyi Cantongqi, Baopuzi Neipian and Tiangong Kaiwu. The source records undergo OCR correction, character normalization, terminology identification and passage segmentation. Each passage retains its document identifier, textual location, original wording and normalized terminology. For example, the “Jindan” chapter of Baopuzi Neipian (juan 4) describes a transformation involving dansha (丹砂) and shuiyin (水銀). The passage provides a documented relationship between two historical substance terms, which can be associated with their modern mineralogical identities through separately attributed scholarly knowledge. The resulting source-linked passages serve as retrieval units for the knowledge graph.

Visual sources comprise historical apparatus illustrations, furnace diagrams, photographs and spatial scans. Provenance is recorded for each image and 3D asset, and shared object identifiers link them to historical terms and apparatus entries. This process produces standardized multimodal knowledge units with identifiable sources, aligned terminology and corresponding objects, providing the input for semantic extraction and knowledge modeling (Fig. 2).

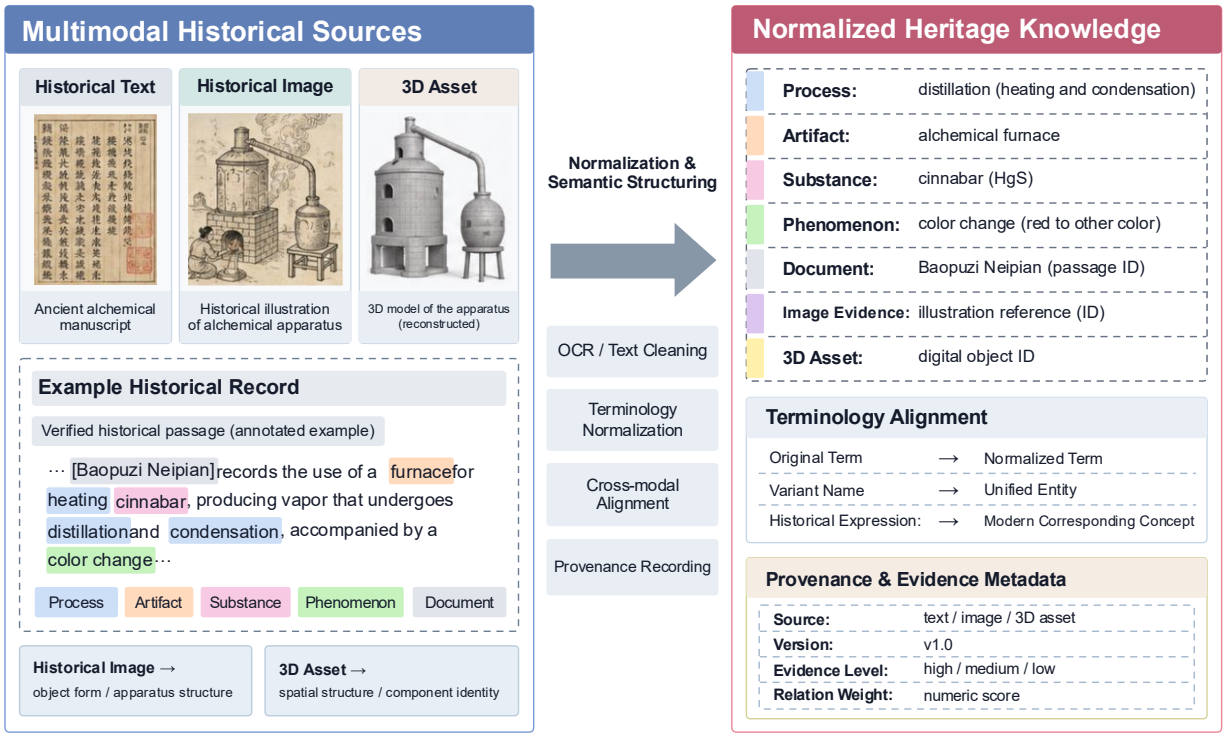


Fig. 2 Normalization and semantic structuring of multimodal historical data. The schematic example illustrates the semantic mapping of a paraphrased historical record.

2.2.2 Collaborative multi-agent semantic processing

A collaborative semantic processing network comprising five agent types operates on the standardized data. It performs entity and relationship extraction, terminology alignment, knowledge merging, conflict checking and semantic organization for subsequent graph retrieval. A scheduling module invokes the agents in task order, and the agents exchange structured results through a common JSON intermediate format (Fig. 3).

The Ingestion and Normalization Agent (INA) executes data ingestion and maintains provenance and metadata records to keep normalization results traceable.

The Entity and Relation Extraction Agent (EA) identifies candidate Process, Artifact, Substance, Principle and Document entities and extracts their relationships from normalized passages. Each extracted record contains the entity label, normalized term, source passage identifier, candidate relationship and confidence score. The extraction results are organized into structured entity–relation records for terminology alignment, evidence checking and subsequent graph construction.

The Alignment and Merging Agent (AMA) aligns alternative names, merges synonymous entities and maps historical terms to modern concepts. References to the same object across documents and modalities are

assigned a shared identifier.

The Conflict Resolution and Auditing Agent (CRA) evaluates extracted relationships against their source passages, normalized terminology and associated evidence metadata. Cross-source inconsistencies are distinguished as terminology differences, relationship discrepancies or conflicting interpretations. The affected records are returned to EA and AMA for targeted re-extraction or entity alignment, while the original source references and conflict annotations are preserved. This procedure maintains the evidential context of competing documentary statements during knowledge graph construction.

The Graph Planning and Retrieval Agent (GPA) organizes validated entities and relationships into process-linked retrieval targets. During knowledge preparation, it associates graph paths with the relevant Process and Artifact entities. During XR interaction, GPA uses the active entities and question type to select passage retrieval, graph traversal or both. It examines the returned passages and paths for documentary support and relevant relationships, and adjusts the retrieval focus when further evidence is required. The three retrieval configurations use the same prepared source materials.

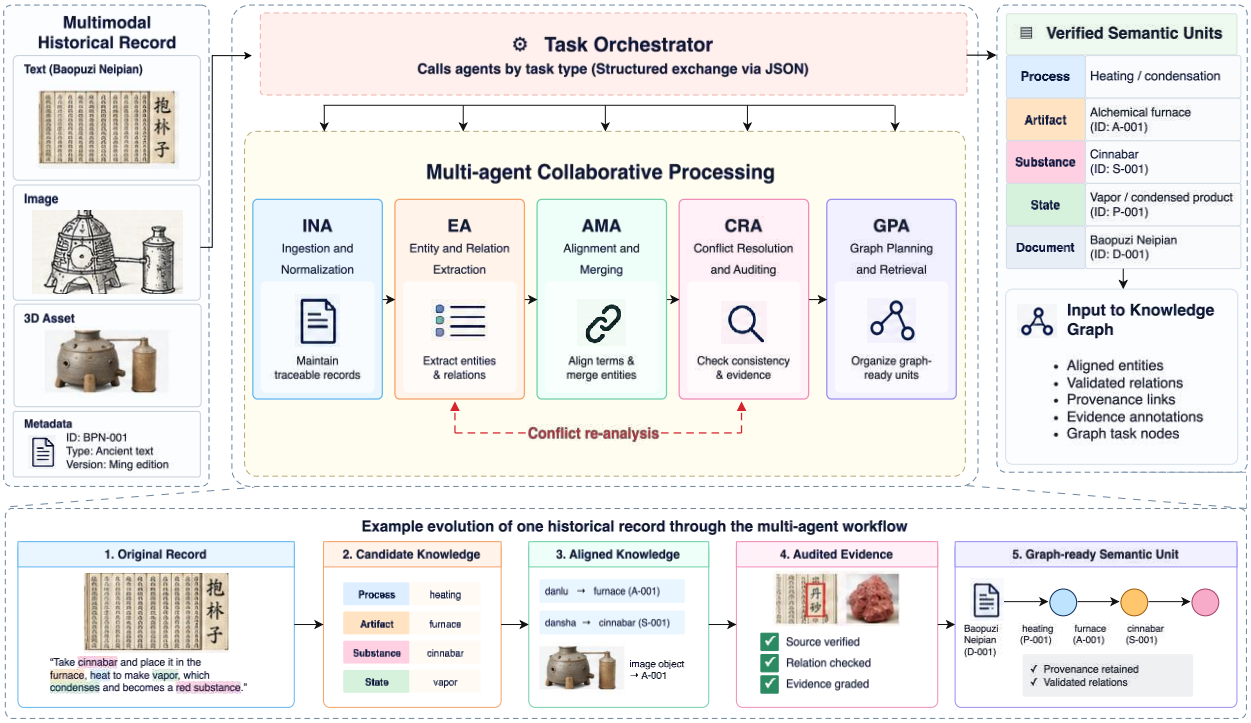


Fig. 3 Collaborative multi-agent semantic processing workflow

2.2.3 Construction of the alchemical heritage knowledge graph

Following multi-agent semantic processing, we constructed a domain knowledge graph that organizes historical records, technical processes, apparatus structures, material states and modern scientific explanations. The implemented graph comprises 12 node labels and 26 relationship types, with five core semantic categories emphasized for contextual interpretation: Process, Artifact, Substance, Principle and Document. These represent process steps, apparatus and components, materials and material states, historical or modern explanatory principles, and source documents, respectively (Fig. 4).

Entities are connected through USES, REQUIRES, PRODUCES, DOCUMENTED_IN, MAPS_TO,

EXPLAINED_BY and SUPPORTED_BY relationships. Process nodes link apparatus and materials through operational relationships and historical passages through DOCUMENTED_IN. In the Baopuzi Neipian example, the recorded material transformation is represented by a Process entity connected to Substance entities for dansha and shuiyin, with a DOCUMENTED_IN link to juan 4, “Jindan”. The historical substance names are retained as source terminology, while MAPS_TO and EXPLAINED_BY provide separately attributed connections to modern substance descriptions and scientific concepts. This representation allows documentary relationships and scientific explanations to be retrieved together while preserving their distinct evidential roles.

The knowledge graph is implemented in Neo4j as a property graph comprising 12 node labels, 26 relationship types and associated provenance metadata. Five core semantic categories and seven key relationship types are emphasized for contextual retrieval. Each entity is assigned a stable identifier, while node properties preserve original terminology, normalized names, aliases and semantic attributes. Document entities retain source and edition information, and source-linked relationships associate technical statements with their documentary evidence and evidential support. Process, Artifact and Substance entities are connected through operational relationships, whereas explanatory relationships link historical practices and material phenomena to Principle entities. This organization supports Cypher-based traversal from an active XR object or process node to the corresponding documentary sources and modern scientific concepts, providing structured evidence for contextual interpretation.

For retrieval and interpretation, source-linked statements are distinguished from modern explanatory links. A DOCUMENTED_IN relation identifies the historical record associated with a technical description, whereas EXPLAINED_BY and SUPPORTED_BY associate interpretations with their stated evidence. This distinction enables source attributions to be retained when historical terminology is mapped to contemporary scientific vocabulary.

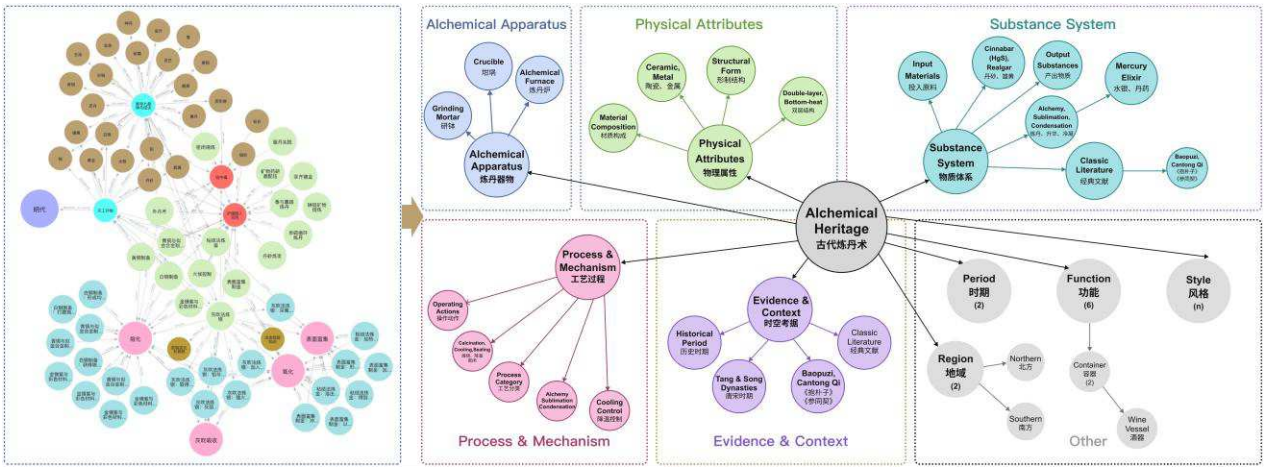


Fig. 4 Provenance-aware knowledge graph of ancient Chinese alchemical heritage.

2.3 Intelligent generation layer

The intelligent generation layer contains two components: 3D asset construction with semantic binding, and

contextual explanation using Agentic GraphRAG. The first assigns graph identifiers to reconstructed apparatus and their interactive parts. The second uses the current XR state to retrieve relevant documents and graph relationships, which are then used to generate explanations for the active task.

2.3.1 3D asset construction and semantic binding

The 3D apparatus assets translate structures and processes described in alchemical texts into spatial objects for XR interaction. Core models were developed from physical records, photogrammetry, historical images and textual descriptions, with structural information cross-referenced across sources.

Three-dimensional reconstruction draws on available documentary and material evidence, with source attribution distinguishing documented structural features from interpretive reconstruction. Historical studies of alchemical apparatus provide contextual evidence for identifying component structures and technical functions [45]. Each interactive 3D object is assigned a semantic identifier that connects its component hierarchy and interaction states with corresponding Artifact, Process and Substance entities in the knowledge graph. This association enables virtual apparatus components to function as identifiable semantic objects during embodied interaction (Fig. 5).

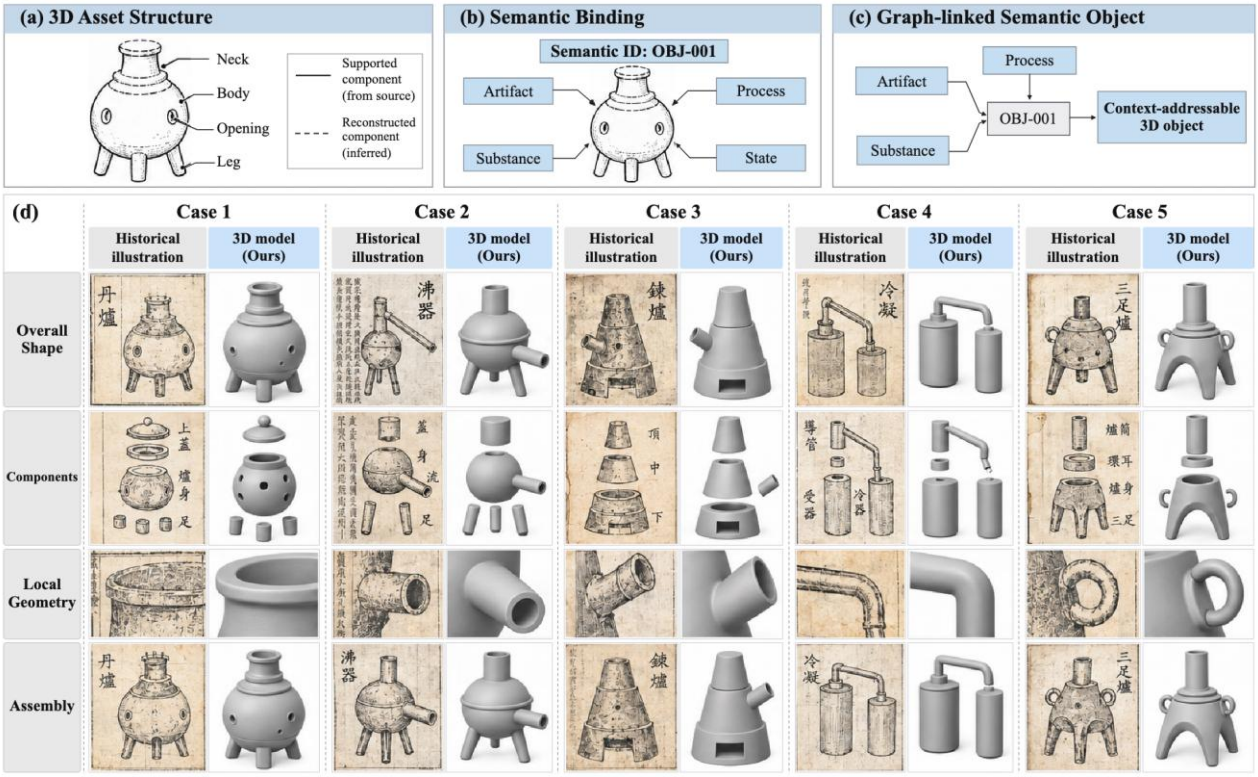


Fig. 5 Evidence-linked reconstruction and semantic binding of alchemical apparatus. (a) Annotated 3D asset structure; (b) object-level semantic binding; (c) graph-linked object representation; and (d) representative comparisons between historical references and reconstructed models.

2.3.2 Contextual interpretation with Agentic GraphRAG

Context routing and hybrid retrieval

Agentic GraphRAG receives the current semantic state and the user's question from the interaction layer to identify the required interpretation. Inputs comprise the active process node, object identifier, material state, recent action and explicit question. Interpretation requests are classified into six categories: terminology clarification, apparatus function, procedural reasoning, material-state change, historical – modern scientific relationships and cross-source comparison. Each category defines a corresponding retrieval focus: terminology requests prioritize normalized names and MAPS_TO relationships; apparatus and procedural requests focus on USES, REQUIRES and PRODUCES relationships; material-state and scientific questions access EXPLAINED_BY connections; and cross-source comparisons retrieve DOCUMENTED_IN and SUPPORTED_BY evidence paths.

Contextual retrieval combines dense passage retrieval with relationship-constrained graph traversal according to the current interaction state and interpretation request. FAISS [46] provides similarity-based candidate retrieval over textual passages and queries represented by the multilingual BGE-M3 model [47] as 1024-dimensional dense embeddings. Cosine similarity between normalized query and passage embeddings provides a common relevance measure for ranking documentary evidence. Graph retrieval uses parameterized Cypher queries rooted in the active Process and Artifact identifiers to retrieve relevant operational relationships, explanatory concepts and source-linked paths in Neo4j. The compact Qwen3.5-9B model serves as the language-model backbone for query planning, evidence integration and contextual explanation generation. Retrieved passages and graph paths are assembled with their associated source identifiers, allowing contextual explanations to be linked to the underlying documentary and semantic evidence.

Embodied State as Query represents the retrieval context as $z_t = [S_t, q_t]$, where $S_t = (p_t, o_t, a_t, m_t, h_t)$ encodes the current embodied state and q_t denotes the user's explicit question. Within S_t, p_t identifies the current Process node, o_t the active object, a_t the latest interaction event, m_t the simulated material state and h_t the sequence of recent interaction events. The object identifier links the selected XR asset to its Artifact entity, while the process and material-state fields identify relevant procedural and Substance relationships. The interaction history provides context for interpreting the current event in relation to preceding operations. Together, these fields determine the retrieval scope by associating the user's question with the active process, object and material condition (Fig. 6).

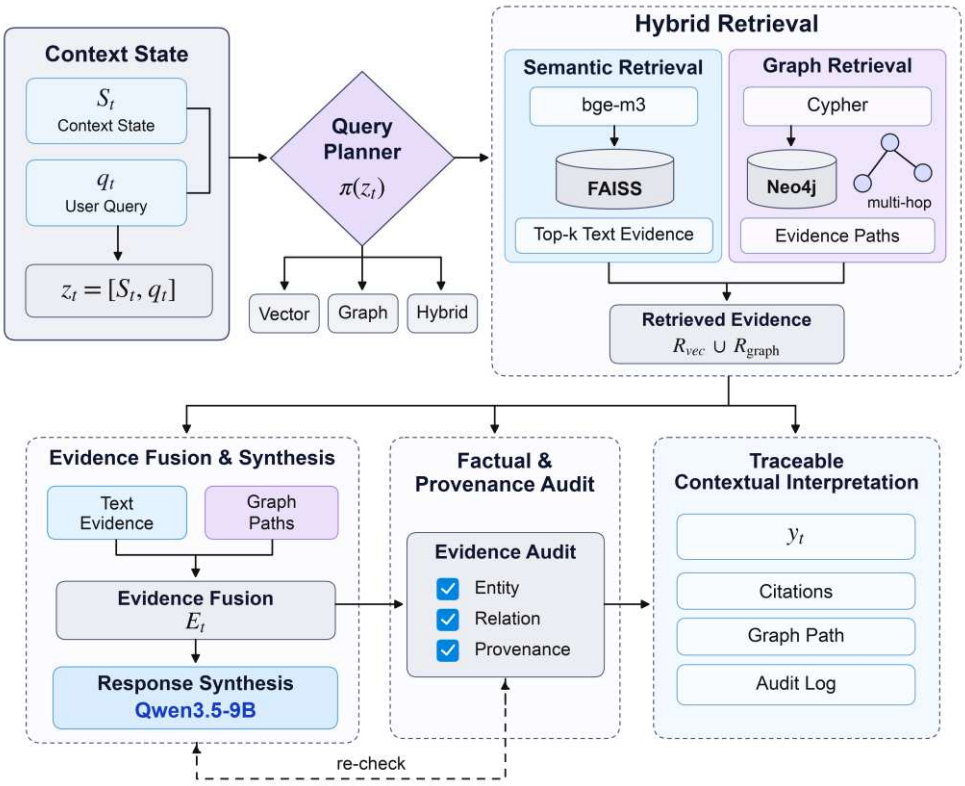


Fig. 6 State-conditioned Agentic GraphRAG for contextual interpretation

Response generation and evidence auditing

Retrieved passages and graph paths provide the source material for explanations presented during XR interaction. The response is organized around historical evidence, technical practice, observed material phenomena and modern scientific interpretation. The Baopuzi Neipian case illustrates this connection: a question about the simulated relationship between dansha and shuiyin can be linked to the "Jindan" passage through DOCUMENTED_IN, while the corresponding Substance and Principle entities provide access to modern mineralogical explanations. The resulting interpretation presents the historical account and its scientific explanation as separately attributed information associated with the current interaction.

Generated responses undergo factual and provenance auditing against the retrieved evidence context. The auditing procedure evaluates three evidence correspondences: entity consistency between generated terminology and knowledge graph records, citation consistency between explanatory statements and retrieved documentary passages, and relational consistency between asserted technical or scientific connections and supporting graph paths. Documentary statements retain their source and edition references, while modern scientific explanations are associated with the corresponding explanatory evidence. When historical sources contain differing interpretations, the response preserves their respective documentary attributions and evidence annotations. The audited relationships provide the evidential basis for the contextual explanation presented in XR.

Each explanation is stored with a retrieval trace containing the interaction state, question category, tool calls, query parameters and returned evidence. The record preserves the order of tool calls, the passage and graph-

path identifiers returned at each step, and the evidence-auditing decision associated with the response. It also links the final explanation to the relevant source references and XR interaction event. These fields allow the retrieval sequence to be reconstructed from an interaction event to the evidence presented in XR (Fig. 7).

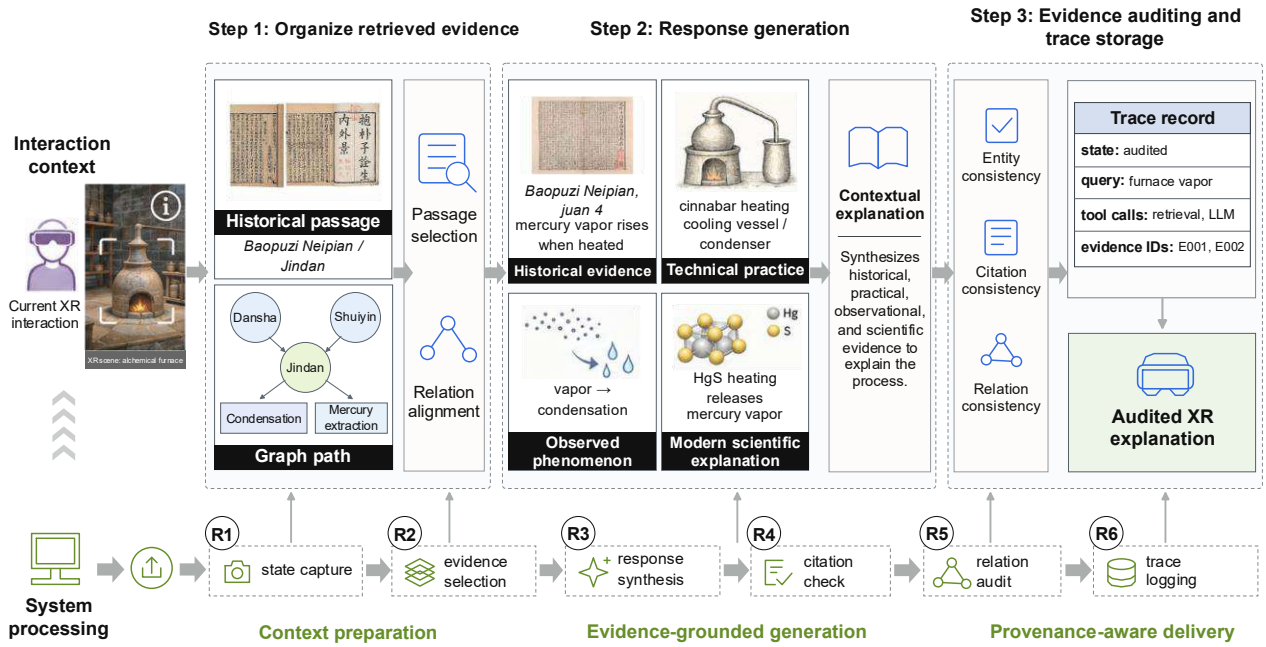


Fig. 7 Evidence-grounded response generation and provenance auditing for contextual interpretation in XR. The figure illustrates three stages: (1) organization of retrieved historical passages and knowledge graph paths, (2) generation of contextual explanations integrating historical evidence and modern scientific interpretations, and (3) entity, citation and relational consistency auditing with retrieval trace storage.

2.4 Interactive presentation layer

The XR interface displays reconstructed apparatus and records the user's current object, process stage, action, parameter settings and material state. These fields are passed to Agentic GraphRAG as retrieval context. When an interaction event updates the state, the interface presents the corresponding process information and explanation alongside the active object.

2.4.1 Historical process contexts and interaction tasks

The immersive XR environment was developed in Unity, with sequential tasks based on representative historical alchemical processes. Each task corresponds to process steps, apparatus configurations and material states in the knowledge graph, translating procedural descriptions in historical texts into executable spatial actions.

The representative XR experience comprises four successive interaction stages: process orientation, virtual apparatus assembly, parameter adjustment and material-state observation (Fig.8). During process orientation, the interface identifies the current task and its associated apparatus. During assembly, users grasp and position reconstructed components through spatial hand interactions, with component identities linked to the corresponding Artifact entities. The parameter interface records adjustments to simulated process conditions, while the observation stage presents changes in virtual material appearance and process-state feedback. Each

stage is associated with a Process node that specifies the relevant apparatus entities, interaction objectives and contextual knowledge accessible during the task.

Particle effects and changes in the virtual material's appearance provide visual feedback during the simulated process. Task progression follows the defined sequence of apparatus assembly, parameter adjustment and observation. Completion of each interaction stage updates the process state used for subsequent knowledge retrieval.

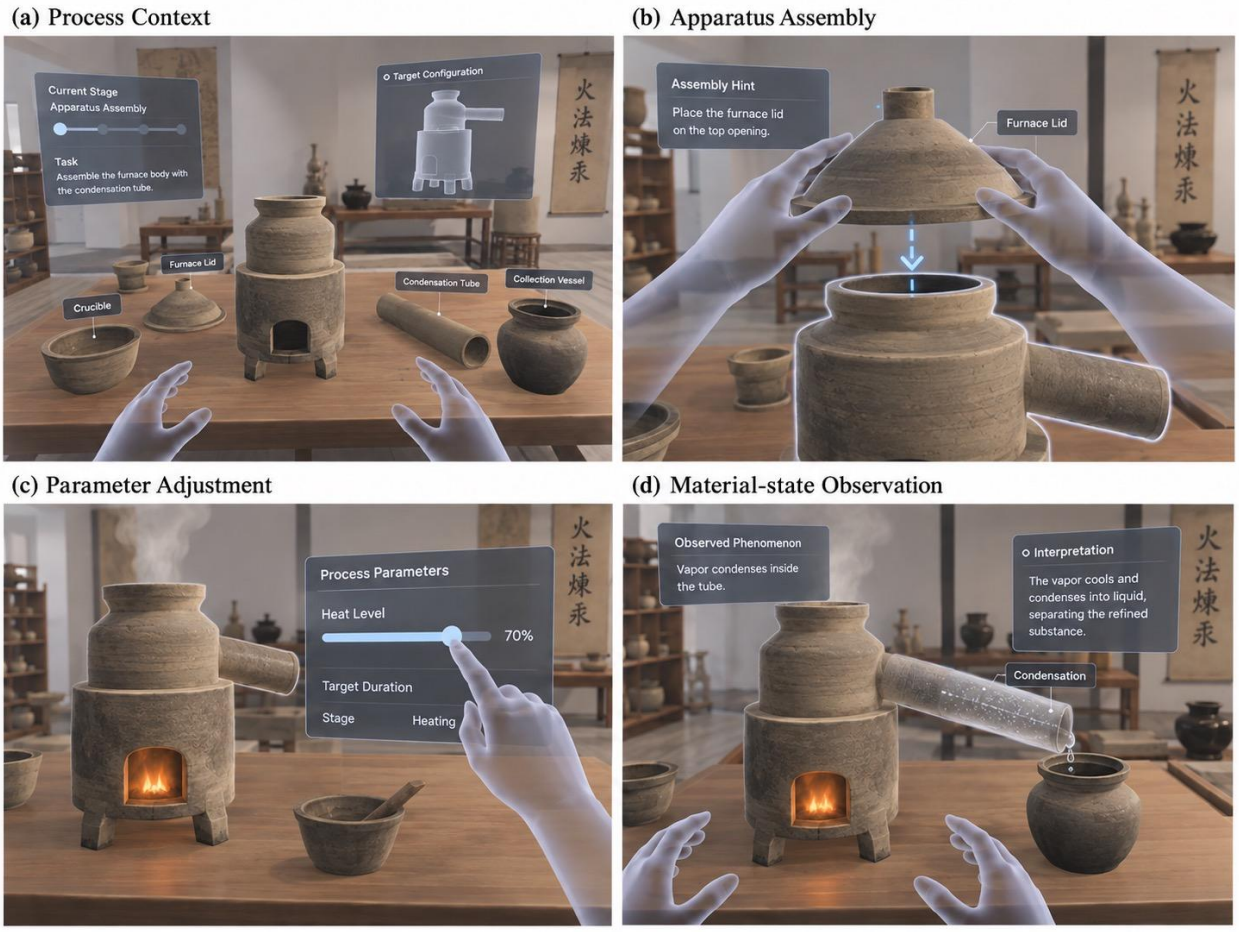
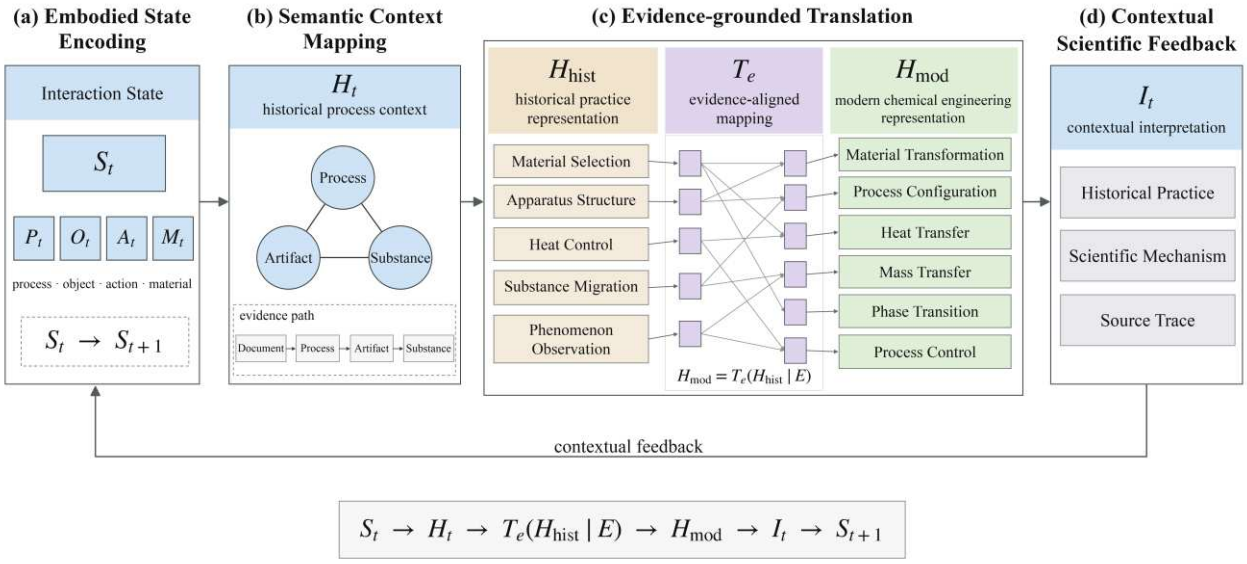


Fig. 8 Representative XR scenes for interaction with historical alchemical processes

2.4.2 Embodied-state mapping and contextual feedback

During interaction, the system records the current process stage, object, action, material state and recent operations as a semantic state. Events such as grasping, spatial movement, assembly sequence and parameter adjustment are first matched to the semantic identifiers of 3D objects. They are then mapped to Process, Artifact and Substance entities and their relationships to identify the user's current process context.

Selecting an apparatus component updates the active object identifier; completing an assembly step updates the process node; and changing a simulated parameter updates the relevant process and material-state fields. The resulting state record is used to select documents and graph relationships for the next explanation (Fig. 9).



$$S_t \rightarrow H_t \rightarrow T_e(H_{\text{hist}} \mid E) \rightarrow H_{\text{mod}} \rightarrow I_t \rightarrow S_{t+1}$$

Fig. 9 Embodied-state-driven evidence-grounded translation and contextual scientific feedback.

Contextual feedback is triggered by recognized interaction events, process-state transitions and explicit user questions. Apparatus selection establishes the active Artifact context, assembly completion updates the corresponding Process state, and simulated parameter adjustments or material-state transitions update the associated procedural and Substance information. These events provide the structured context for Agentic GraphRAG to select relevant documentary passages and explanatory relationships. During assembly, feedback focuses on apparatus functions and procedural relationships; during material-state observation, explanations connect the represented phenomenon with its historical description and relevant modern scientific concepts. Retrieved explanations are presented through the XR dialogue interface together with their associated source information.

The XR interface displays generated text, state prompts and visual feedback during interaction. Explanations retain the associated process identifier and documentary references. Depending on the current task, the content describes the historical operation, the function of the apparatus or a modern scientific account of the simulated material phenomenon (Fig. 10).

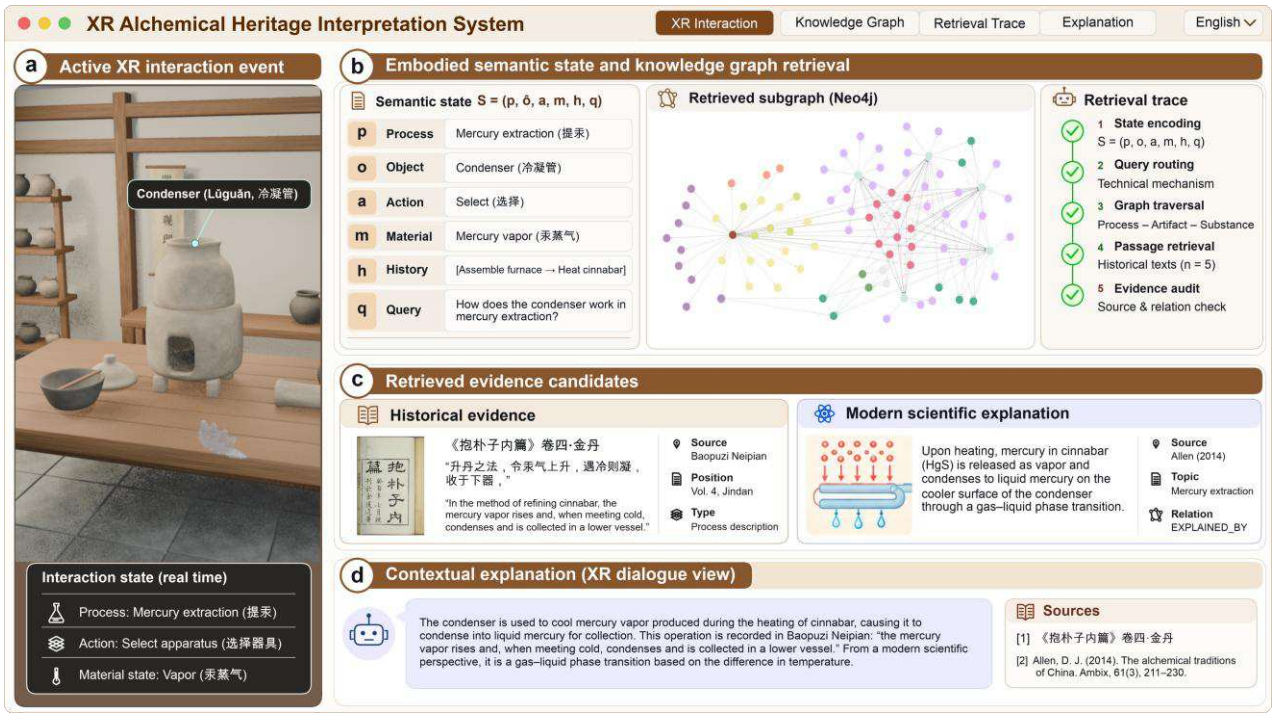


Fig.10 Example of embodied-state-conditioned evidence retrieval during XR interaction.
(a) Active XR interaction event; (b) semantic state and knowledge graph retrieval; (c) retrieved historical and scientific evidence; and (d) contextual explanation displayed in XR.

2.5 Offline evaluation

2.5.1 Compared retrieval conditions and shared settings

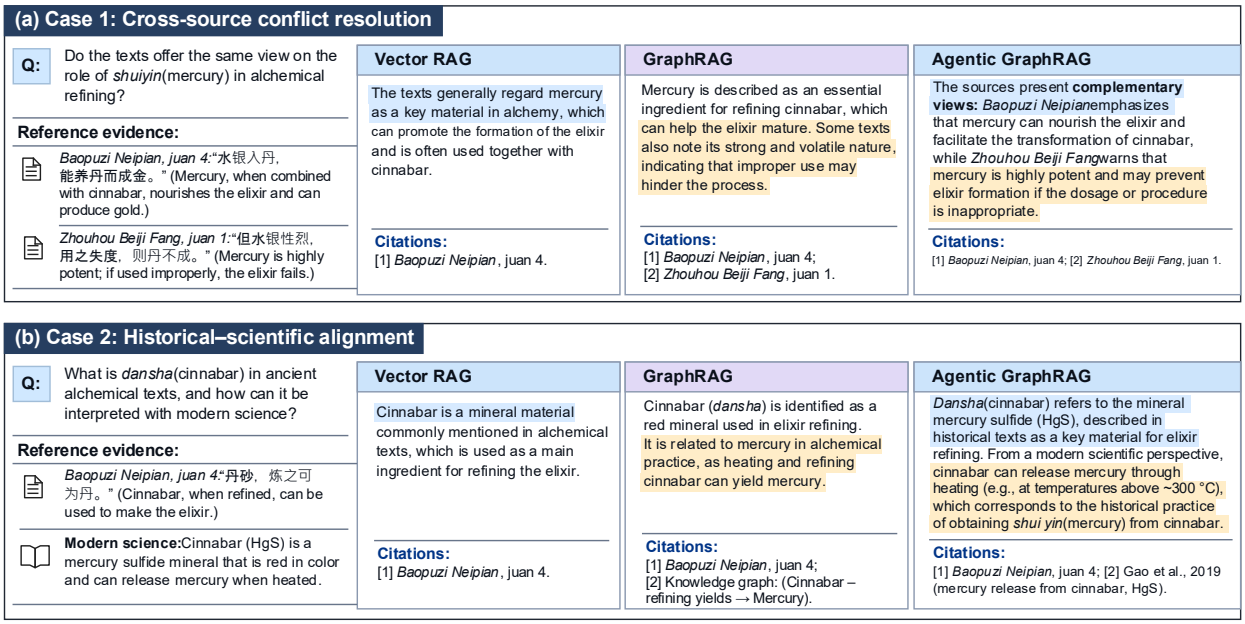
The offline experiment compared three retrieval configurations using the same historical corpus, evaluation questions and reference materials. Vector RAG ranked textual passages by embedding similarity, while GraphRAG additionally retrieved entity relationships and graph paths. Agentic GraphRAG combined these retrieval resources with question-type routing, tool selection, iterative retrieval and evidence auditing. The embedding model, language-model backbone, response-length limit and citation requirements were held constant across conditions. The comparison assessed the interpretation quality produced by each complete retrieval configuration under a common set of questions.

2.5.2 Evaluation questions and reference evidence

The evaluation question set was developed from historical alchemical texts, scholarly studies and the domain knowledge graph. The questions were organized into six interpretation categories: terminology equivalence, apparatus functions, procedural dependencies, material-state changes, cross-source differences and historical–modern scientific relationships. Reference evidence comprised relevant documentary passages, their associated graph entities and relationships, and corresponding modern scientific literature. Each question was evaluated against the same reference evidence across all three retrieval configurations. This organization supports category-specific assessment of terminology interpretation, procedural reasoning, provenance attribution and scientific explanation.

2.5.3 Interpretation-quality metrics

Retrieval and interpretation quality was assessed using four complementary metrics, each defined by a specific evaluation unit and corresponding evidence criterion. Conflict resolution accuracy used source-conflict cases as evaluation units and assessed agreement with reference interpretations. Answer factual consistency used factual statements and assessed their support from documentary passages or knowledge graph relationships. Citation traceability completeness used evidence-dependent statements and assessed their correspondence with identifiable source references or graph paths. Modern scientific alignment accuracy used scientific explanations and assessed their consistency with the corresponding reference knowledge. Each metric was calculated as the number of evaluation units meeting its criterion divided by the total number of eligible units, yielding a score between 0 and 1 (Fig. 11).



Metric	Evaluation focus	Case 1			Case 2		
		Vector	Graph	Agentic	Vector	Graph	Agentic
Conflict resolution accuracy	Whether competing or complementary views are correctly identified and reconciled.	×	✓	✓	—	—	—
Factual consistency	Whether key statements are supported by the reference evidence.	✓	✓	✓	✓	✓	✓
Citation traceability completeness	Whether statements are properly attributed to identifiable sources.	×	✓	✓	×	✓	✓
Modern scientific alignment accuracy	Whether the historical description is correctly interpreted with modern scientific knowledge.	—	—	—	×	✓	✓

Fig. 11 Representative cases of evidence-based retrieval and interpretation quality evaluation. (a) Cross-source conflict resolution and (b) historical–scientific alignment, comparing the responses of Vector RAG, GraphRAG and Agentic GraphRAG against reference evidence. (c) Case-level assessment using four interpretation-quality metrics.

2.6 User study

The user study examined heritage knowledge, cognitive workload, user experience and exploratory behavior during a simulated alchemical process involving cinnabar. Sixty-six participants were assigned to three conditions (n = 22 per group). Group A used the XR system with state-conditioned explanations generated

by Agentic GraphRAG, while Group B performed the same XR tasks with scripted feedback presented at predefined stages. Group C accessed the same core content through Web-based texts, images, flowcharts and static 3D models. The A – B comparison examined differences between feedback strategies under common XR tasks, while the Web-based condition provided a reference for learning through a different presentation format (Fig. 12).

H1: Knowledge gains. The complete XR system may support greater conceptual and procedural knowledge gains than a Web-based presentation of comparable content; the additional contribution of dynamic feedback can be examined by comparing the two XR conditions.

H2: Workload. Group differences in perceived workload are assessed to determine whether the immersive and dynamic-feedback conditions impose additional overall or dimension-specific demands.

H3: Interpretive experience and exploration. Embodied-state-driven explanations may be associated with greater exploration of historical evidence and scientific mechanisms, perceived explanation credibility, source clarity and user experience.

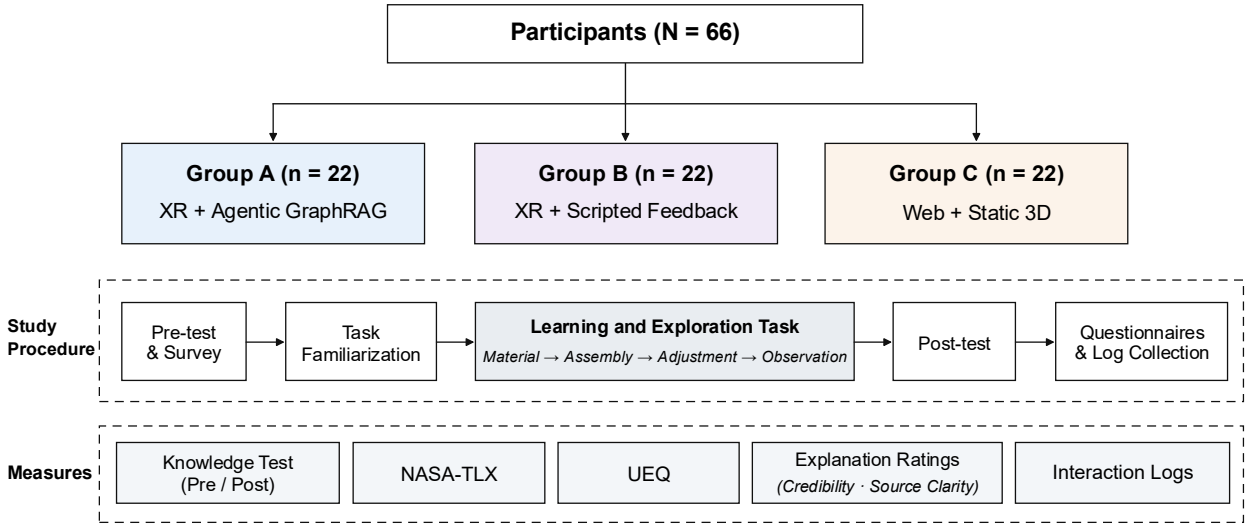


Fig. 12 User study design and experimental procedure

2.6.1 Experimental procedure

Each participant completed an individual session in a controlled human – computer interaction laboratory. Sessions lasted approximately 60 – 70 min and included participant information collection, pretesting, equipment and gesture familiarization, the assigned learning intervention, posttesting and subjective evaluation. Learning activities covered material identification, virtual apparatus assembly, simulated parameter adjustment, material-state observation and knowledge exploration. Behavioral measures included participant-initiated information requests, assembly or parameter errors, task completion time and the proportion of mechanism-oriented questions. Participant-initiated requests referred to explicit inquiries, distinguishing them from feedback triggered by XR state transitions. Mechanism-oriented questions concerned the reasons for technical operations, relationships between process conditions and material phenomena, and associated scientific explanations. Their proportion was calculated relative to each participant's recorded information requests.

2.6.2 Measures

Heritage knowledge was assessed before and after the intervention across three dimensions: factual knowledge concerning historical terminology, materials and apparatus (0 – 6 points); conceptual knowledge concerning technical functions and relationships between material phenomena and scientific explanations (0 – 7 points); and procedural knowledge concerning process sequences, apparatus relationships and operational dependencies (0 – 7 points). The total score ranged from 0 to 20 points. Knowledge gain was calculated separately for each dimension and for the total score as the posttest score minus the corresponding pretest score.

Cognitive workload was assessed using the unweighted NASA Task Load Index (Raw NASA-TLX) [48], covering mental demand, physical demand, temporal demand, perceived performance, effort and frustration. The overall score was the arithmetic mean of the six dimensions, with lower scores indicating lower subjective workload.

User experience was assessed using the standard 26-item User Experience Questionnaire (UEQ) [49], covering attractiveness, efficiency, perspicuity, dependability, stimulation and novelty. Each dimension was scored from −3 to +3. The Mean UEQ score reported in Table 4 summarizes the six dimensions using an equal-weight arithmetic mean. Two supplementary seven-point items assessed perceived explanation credibility and source clarity. Explanation credibility measured participants' ratings of the reliability of the feedback, while source clarity assessed how clearly documentary sources were communicated.

2.6.3 Data processing and statistical analysis

Statistical analyses were conducted at the participant level. Knowledge gain was calculated as the posttest score minus the pretest score for each knowledge domain and the total score. Total knowledge gain summarized overall learning performance, while factual, conceptual and procedural gains were examined separately. One-way analyses of variance (ANOVAs) were used for the reported three-group comparisons, with F statistics, two-sided p values and partial η^2 as the effect-size measure. Significant omnibus effects were followed by Tukey's honestly significant difference (HSD) tests. Approximate 95% Tukey simultaneous confidence intervals for selected contrasts were calculated from the reported group means, standard deviations and sample sizes using pooled within-group variance and the studentized-range distribution. Assembly or parameter errors were compared between the two XR groups using an independent-samples t test, with Cohen's d as the effect-size measure [50]. The significance threshold was $\alpha = 0.05$. Results from separate knowledge, workload, user-experience and behavioral measures were interpreted within their respective outcome domains.

2.6.4 Ethics approval and informed consent

This study was conducted in accordance with the institutional research guidelines of East China University of Science and Technology. Following institutional review, formal ethics committee approval was deemed unnecessary for this non-clinical educational study. All procedures involving human participants were conducted in accordance with the principles of the Declaration of Helsinki. All participants were informed of the study purpose and procedures and provided written informed consent prior to participation.

3. Results

3.1 Offline retrieval and interpretation quality

Agentic GraphRAG obtained the highest score on each of the four offline metrics (Table 1; Fig. 13). Conflict resolution accuracy was 0.848, compared with 0.535 for GraphRAG and 0.365 for Vector RAG. Factual consistency reached 0.883, while citation traceability and modern scientific alignment reached 0.857 and 0.866, respectively. Compared with GraphRAG, the score differences were 0.313 for conflict resolution, 0.080 for factual consistency, 0.202 for citation traceability and 0.380 for modern scientific alignment. The largest differences concerned the handling of competing evidence and the interpretation of material phenomena using modern scientific knowledge.

Table 1 Performance comparison of retrieval and interpretation methods

Metric	Vector RAG	GraphRAG	Agentic GraphRAG
Conflict resolution accuracy	0.365	0.535	0.848
Answer factual consistency	0.720	0.803	0.883
Citation traceability completeness	0.542	0.655	0.857
Modern scientific alignment accuracy	0.292	0.486	0.866

Note: Scores range from 0 to 1; higher values indicate better performance.

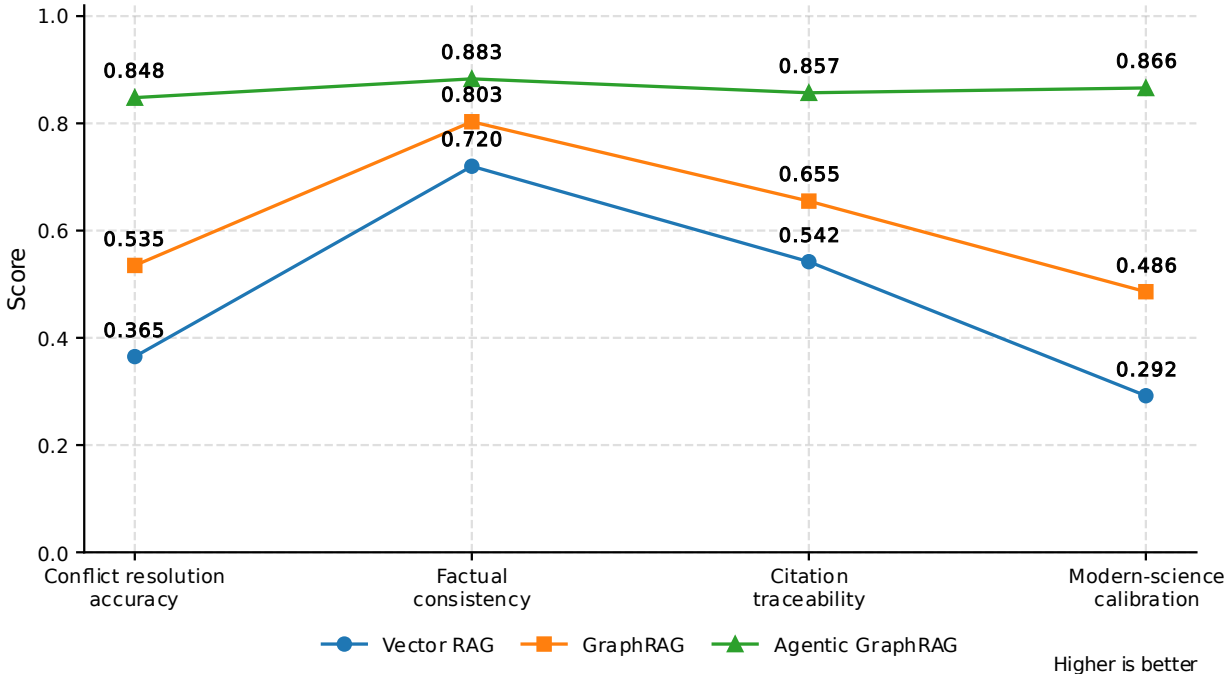


Fig. 13 Offline evaluation of retrieval and interpretation quality

3.2 Knowledge gains

Knowledge gains were compared across factual, conceptual, procedural and overall knowledge dimensions

(Table 2; Fig. 14). Significant group effects were observed for conceptual knowledge, $F(2, 63) = 3.73$, $p = 0.029$, partial $\eta^2 = 0.106$; procedural knowledge, $F(2, 63) = 4.47$, $p = 0.015$, partial $\eta^2 = 0.124$; and total knowledge gain, $F(2, 63) = 7.42$, $p = 0.001$, partial $\eta^2 = 0.191$. Tukey HSD comparisons identified significant differences between Groups A and C in these three dimensions. Factual knowledge gains showed an overall group comparison of $F(2, 63) = 0.88$, $p = 0.421$.

Table 2 Knowledge gains across experimental conditions

Knowledge domain	Group A M (SD)	Group B M (SD)	Group C M (SD)	F(2,63)	p value	Partial η^2
Factual knowledge (0–6)	1.36(1.40)	1.05(1.05)	0.95(0.65)	0.88	0.421	0.027
Conceptual knowledge (0–7)	1.59(1.37)	1.36(1.14)	0.68(0.89)	3.73	0.029	0.106
Procedural knowledge (0–7)	2.00(1.69)	1.55(1.34)	0.77(1.02)	4.47	0.015	0.124
Total knowledge gain (0–20)	4.95(2.57)	3.95(2.34)	2.41(1.59)	7.42	0.001	0.191

Note: *p < 0.05, **p < 0.01, ***p < 0.001. Partial η^2 denotes effect size.

Mean total knowledge gains were 4.95 ± 2.57, 3.95 ± 2.34 and 2.41 ± 1.59 in Groups A, B and C, respectively. Group A achieved a 2.54-point greater total knowledge gain than Group C, with an approximate 95% Tukey simultaneous confidence interval of 0.94–4.14 points. The corresponding A–C differences were 0.91 points for conceptual knowledge (95% CI 0.08–1.74) and 1.23 points for procedural knowledge (95% CI 0.23–2.23). Within the shared XR environment, the A–B difference in total knowledge gain was 1.00 point (95% CI −0.60 to 2.60). These comparisons identify the clearest statistically supported learning differences between the adaptive XR and Web-based conditions.

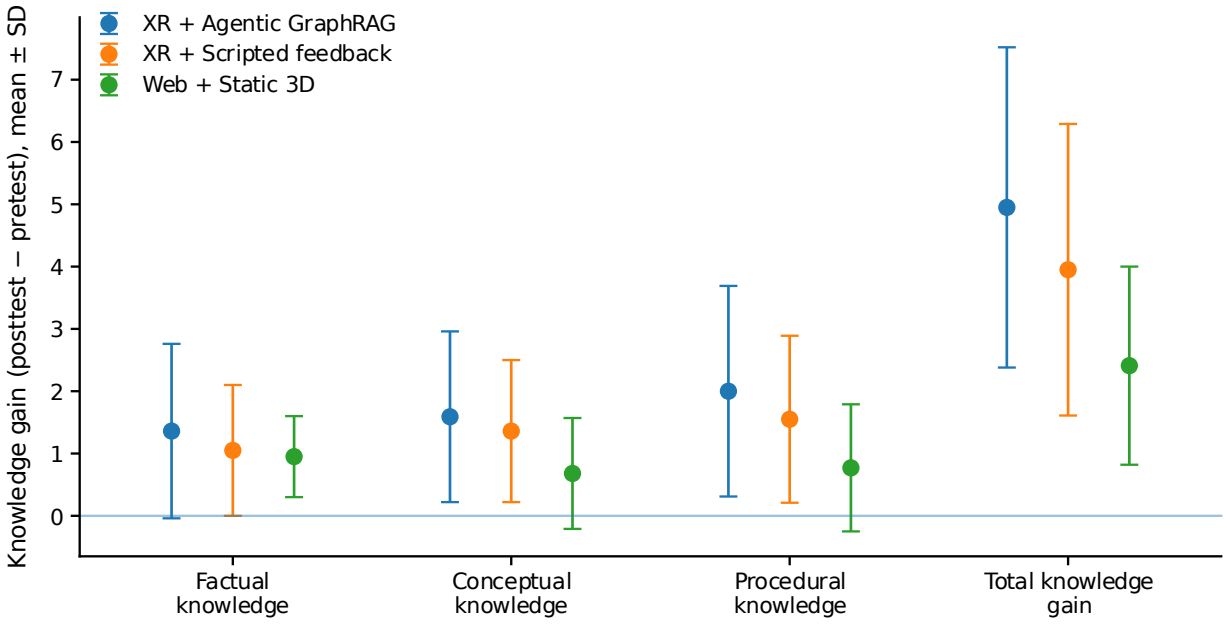


Fig. 14 Factual, conceptual, procedural and total knowledge gains across the three groups (M ± SD)

3.3 Cognitive workload

Overall Raw NASA-TLX scores were 43.23 ± 6.22, 46.39 ± 8.47 and 48.30 ± 6.51 in Groups A, B and C, respectively, $F(2, 63) = 2.83$, $p = 0.067$, partial $\eta^2 = 0.082$ (Table 3). Significant group effects were observed for mental demand ($p = 0.020$), physical demand ($p < 0.001$), perceived performance ($p < 0.001$) and frustration ($p < 0.001$). Group A reported the lowest mean mental demand (43.27 ± 11.28) and frustration (37.18 ± 9.26), while physical demand was higher in both XR conditions than in the Web-based condition.

Table 3 NASA-TLX cognitive load across experimental conditions

NASA-TLX dimension	Group A M (SD)	Group B M (SD)	Group C M (SD)	F(2,63)	p value	Partial η^2
Mental demand	43.27(11.28)	51.56(13.73)	52.44(9.59)	4.15	0.020	0.116
Physical demand	37.07(7.79)	39.70(7.09)	20.32(6.96)	45.70	<0.001	0.592
Temporal demand	46.83(12.53)	50.33(12.38)	53.99(10.78)	1.98	0.146	0.059
Performance (lower is better)	39.47(9.86)	38.67(13.56)	51.45(7.61)	9.98	<0.001	0.241
Effort	55.61(11.67)	52.68(13.86)	62.26(14.28)	2.98	0.058	0.086
Frustration	37.18(9.26)	45.41(8.14)	49.37(9.68)	10.39	<0.001	0.248
Overall workload	43.23(6.22)	46.39(8.47)	48.30(6.51)	2.83	0.067	0.082

3.4 User experience

Group A had the highest mean score on all six UEQ dimensions (Table 4; Fig. 15). Tukey HSD comparisons favored Group A over Groups B and C for attractiveness, perspicuity, dependability, stimulation and novelty; the main efficiency difference was between Groups A and C. In the two supplementary seven-point measures, adaptive XR was rated higher than scripted XR for explanation credibility (5.63 ± 0.70 vs. 5.00 ± 0.68; mean difference 0.63, approximate 95% Tukey CI 0.11 – 1.15) and source clarity (5.51 ± 0.79 vs. 4.68 ± 0.64; mean difference 0.83, approximate 95% Tukey CI 0.29 – 1.37). Both measures showed significant overall group effects ($p < 0.001$).

Table 4 User experience and interpretation quality across experimental conditions

Experience measure	Group A M (SD)	Group B M (SD)	Group C M (SD)	F(2,63)	p value	Partial η^2
Attractiveness	0.93(0.43)	0.41(0.71)	0.20(0.60)	8.82	<0.001	0.219
Efficiency	0.40(0.57)	0.09(0.60)	-0.11(0.58)	4.27	0.018	0.119
Perspicuity	0.76(0.56)	0.03(0.55)	-0.05(0.62)	12.89	<0.001	0.290
Dependability	0.61(0.55)	0.07(0.74)	-0.18(0.61)	8.88	<0.001	0.220
Stimulation	0.88(0.51)	0.21(0.44)	0.09(0.73)	11.93	<0.001	0.275
Novelty	0.97(0.61)	0.10(0.61)	0.02(0.53)	17.71	<0.001	0.360
Mean UEQ score	0.76(0.46)	0.15(0.51)	-0.01(0.51)	14.61	<0.001	0.317
Explanation credibility	5.63(0.70)	5.00(0.68)	4.72(0.77)	9.13	<0.001	0.225
Source clarity	5.51(0.79)	4.68(0.64)	4.48(0.80)	11.83	<0.001	0.273

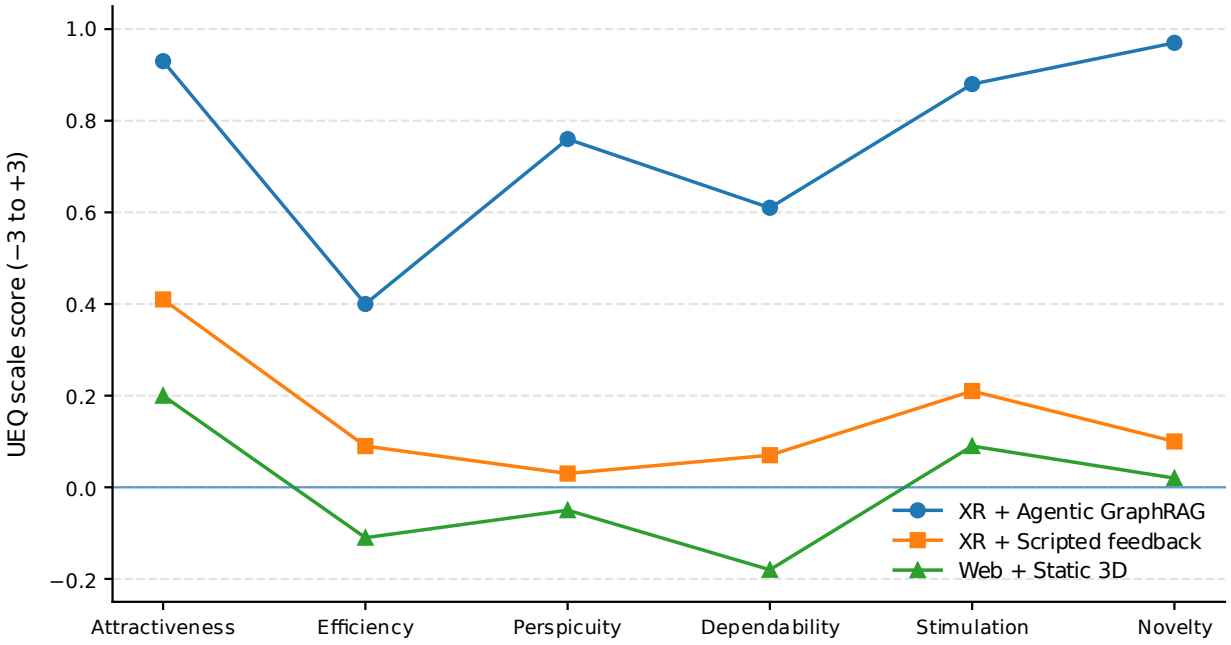


Fig. 15 User experience across the six UEQ dimensions in the three groups

3.5 Exploratory interaction behavior

Participants in Group A made an average of 6.55 ± 3.38 active information requests, compared with 4.18 ± 1.76 in Group B and 3.95 ± 1.89 in Group C (Table 5). The overall group effect was significant, $F(2, 63) = 7.51$, $p = 0.001$, partial $\eta^2 = 0.193$. The difference between Groups A and B was 2.37 requests per participant (approximate 95% Tukey CI 0.59 – 4.15). Mechanism-oriented questions accounted for 0.49 ± 0.19 of requests in Group A and 0.18 ± 0.11 in each comparison group. The A – B difference was 0.31 (95% CI 0.21 – 0.41), equivalent to 31 percentage points. The overall group effect for this measure was $F(2, 63) = 35.14$, $p < 0.001$, partial $\eta^2 = 0.527$.

Table 5 Interaction behavior across experimental conditions

Behavioral measure	Group A M (SD)	Group B M (SD)	Group C M (SD)	Test statistic	p value	Effect size
Active information requests	6.55(3.38)	4.18(1.76)	3.95(1.89)	7.51	0.001	0.193
Assembly or parameter errors (XR only)	2.50(1.63)	4.00(2.31)	Not applicable	t(42)=-2.49	0.017	d=-0.75
Task completion time (min)	26.26(3.84)	28.01(3.87)	22.26(4.62)	11.21	<0.001	0.263
Proportion of mechanism-oriented questions	0.49(0.19)	0.18(0.11)	0.18(0.11)	35.14	<0.001	0.527

Note: Values are presented as mean (SD). Three-group comparisons report F(2, 63) and partial η^2. The XR-only error comparison reports t(42) and Cohen's d, calculated as Group A minus Group B. Negative d values indicate fewer errors in Group A.

Within the shared XR environment, Group A recorded fewer assembly or parameter errors (2.50 ± 1.63) than Group B (4.00 ± 2.31), $t(42) = -2.49$, $p = 0.017$, Cohen's $d = -0.75$ (Fig. 16). Task completion times were 26.26 ± 3.84 min, 28.01 ± 3.87 min and 22.26 ± 4.62 min for Groups A, B and C, respectively, with a significant overall group effect, $F(2, 63) = 11.21$, $p < 0.001$. The shorter completion time in Group C corresponded to its Web-based learning format, which involved different interaction activities from the two XR conditions.

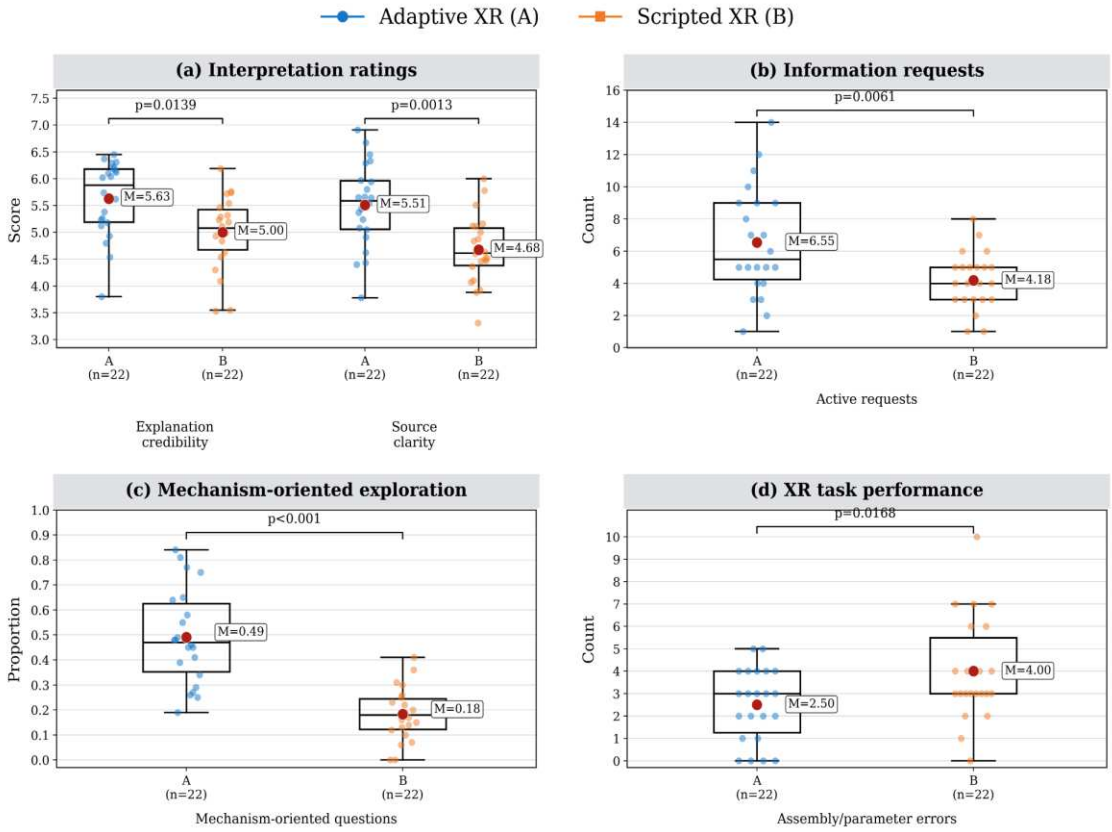


Fig. 16 Adaptive versus scripted XR: interpretation ratings and exploratory behavior. (a) Explanation credibility and source clarity (seven-point measures); (b) active information requests; (c) proportion of mechanism-oriented questions; and (d) assembly or parameter errors in XR. Boxes indicate interquartile ranges, individual points represent participants, and red markers indicate group means (n = 22 per XR condition).

4. Discussion

This study examines how an ongoing XR operation can guide access to historical evidence about ancient Chinese alchemical practices. Its main contribution is the use of embodied interaction states as retrieval conditions. The system maps the current process, object and material state to knowledge graph entities, allowing Agentic GraphRAG to select documentary passages and scientific explanations relevant to the user's activity. This places evidence retrieval within the operation being interpreted, rather than treating knowledge access as a separate question-answering task. The approach is particularly relevant to scientific and

technological heritage, where the meaning of an apparatus often depends on how it was used and what was observed during a procedure.

The offline comparison showed the largest differences in two measures: modern scientific alignment and conflict resolution. Relative to GraphRAG, Agentic GraphRAG scored 0.380 higher on modern scientific alignment and 0.313 higher on conflict resolution, compared with a smaller difference of 0.080 in factual consistency. This pattern suggests that the additional retrieval and auditing mechanisms may be particularly useful when an explanation must connect historical descriptions with scientific concepts or account for differences between sources. A graph path can identify relationships among an apparatus, process and document, while query planning directs retrieval towards the type of evidence required by the current question. Evidence auditing then checks the correspondence between the retrieved sources and the generated explanation. The results are consistent with the combined contribution of these operations, especially for questions involving several forms of historical and scientific evidence.

The user study revealed different patterns for learning outcomes and interpretive experience. Participants in adaptive XR gained an average of 4.95 points on the knowledge test, compared with 2.41 points in the Web-based condition. The differences were most evident in conceptual and procedural knowledge, which concern the relationships among technical functions and process steps. Within the shared XR environment, the clearer contrasts concerned the explanations themselves. Adaptive feedback received higher ratings for credibility (5.63 vs. 5.00) and source clarity (5.51 vs. 4.68), and participants asked more questions about technical mechanisms (0.49 vs. 0.18). These findings point to a specific role for state-conditioned retrieval: it makes documentary and scientific information available at the moment users are examining a related operation, and is associated with more active investigation of how that operation works.

The distinction between historical records and modern scientific explanations is central to interpreting alchemical heritage. A historical account identifies the terminology, apparatus and observations associated with a practice, whereas modern scientific knowledge offers an explanation of the material phenomena involved. The knowledge graph represents these forms of information through different relationships. DOCUMENTED_IN connects a process to its documentary source, while EXPLAINED_BY links the relevant process or material state to a scientific concept. In the XR system, selecting a reconstructed apparatus therefore provides access to both its historical context and its associated scientific explanation, with the two forms of knowledge attributed separately. This structure helps preserve the historical meaning of technical descriptions while making their material implications accessible to present-day learners.

The same retrieval approach may be useful for other heritage practices in which knowledge is tied to a sequence of operations. Traditional crafts, historical manufacturing processes and scientific instruments involve tools, materials and actions whose functions become clearer when considered within a procedure. For such cases, the interaction state can identify the active object and process step, while a knowledge graph connects them to documentary descriptions and explanatory concepts. The relevant transferable principle is the use of the current operation to select evidence, rather than the particular apparatus or historical period represented in the present XR system.

The workload results provide another perspective on embodied heritage interaction. Both XR groups reported greater physical demand than the Web-based group, consistent with the spatial manipulation and assembly tasks required in the immersive environment. At the same time, the adaptive XR group had the lowest mean

mental demand and frustration, while the overall workload comparison yielded $F(2, 63) = 2.83$ and $p = 0.067$. The dimension-level results distinguish the physical effort of manipulating reconstructed apparatus from participants' reported experience of task difficulty and frustration. This distinction is relevant when designing interactive explanations for technical heritage, where bodily participation and knowledge support occur within the same activity.

The study has several limitations. Evaluation focused on one simulated alchemical process and immediate learning outcomes in 66 participants. The adaptive condition combined embodied-state encoding, graph retrieval and evidence auditing, leaving their individual contributions to be examined through component-level experiments. Further evaluation with historical specialists, more detailed question-level assessments and delayed learning tests would provide a stronger basis for assessing source reliability, knowledge retention and applicability across heritage contexts.

Further work could examine how different kinds of historical questions affect retrieval planning. Questions about apparatus functions may be answered through short graph paths, while comparisons between historical accounts may require broader documentary retrieval and additional source checking. Evaluating these retrieval routes separately would help identify suitable agent strategies for different interpretation tasks. Extending the knowledge graph to other historical technologies and incorporating more documentary and archaeological evidence would also provide opportunities to examine state-dependent retrieval across a wider range of technical practices. These developments could inform the design of museum systems that connect physical interaction with historical sources and scientific explanations.

Data availability

The de-identified participant data, knowledge graph data, offline evaluation questions, reference annotations, and analysis materials supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable consent and privacy restrictions.

Code availability

The code used for knowledge graph construction, retrieval, XR integration and statistical analysis is available from the corresponding author upon reasonable request. Third-party models, software and assets remain subject to their respective licences.

Acknowledgments

This study was partly supported by the Research Project of Humanities and Social Sciences of the Ministry of Education (No. 24YJA760013), the National Natural Science Foundation of China (No. 51905175), Shanghai Soft Science Key Project (No. 24692109400), and Xie Youbai Design Science Research Foundation (No. XYB-DS-202301).

Competing interests

The authors declare no competing interests.

Consent for publication

Informed consent was obtained from all subjects involved in the study.

Author contributions

Y.C. contributed to conceptualization, methodology, supervision, and writing—review and editing. T.X. contributed to conceptualization, methodology, data curation, knowledge graph construction, system implementation, formal analysis, visualization, and writing—original draft. Both authors reviewed and approved the final manuscript.

References

[1] Partington, J. R. An Ancient Chinese Treatise on Alchemy. Nature **136**, 287–288 (1935). https://doi.org/10.1038/136287a0

[2] Moran, B. T. Introduction: Focus—Alchemy and the History of Science. Isis **102**, 300–304 (2011). https://doi.org/10.1086/660138

[3] Principe, L. M. Alchemy Restored. Isis **102**, 305–312 (2011). https://doi.org/10.1086/660139

[4] Newman, W. R. What Have We Learned from the Recent Historiography of Alchemy? Isis **102**, 313–321 (2011). https://doi.org/10.1086/660140

[5] Chang, K.-m. Alchemy as Studies of Life and Matter: Reconsidering the Place of Vitalism in Early Modern Chymistry. Isis **102**, 322–329 (2011). https://doi.org/10.1086/660141

[6] Fabbrizzi, L. Communicating about Matter with Symbols: Evolving from Alchemy to Chemistry. J. Chem. Educ. **85**, 1501 (2008). https://doi.org/10.1021/ed085p1501

[7] Marchini, M., Gandolfi, M., Maini, L., Raggetti, L. & Martelli, M. Exploring the ancient chemistry of mercury. Proc. Natl Acad. Sci. USA **119**, e2123171119 (2022). https://doi.org/10.1073/pnas.2123171119

[8] Gliozzo, E. Pigments—Mercury-based red (cinnabar-vermilion) and white (calomel) and their degradation products. Archaeol. Anthropol. Sci. **13**, 210 (2021). https://doi.org/10.1007/s12520-021-01402-4

[9] Sylaiou, S., Liarokapis, F., Kotsakis, K. & Patias, P. Virtual museums, a survey and some issues for consideration. J. Cult. Herit. **10**, 520–528 (2009). https://doi.org/10.1016/j.culher.2009.03.003

[10] Gomes, L., Bellon, O. R. P. & Silva, L. 3D reconstruction methods for digital preservation of cultural heritage: A survey. Pattern Recognit. Lett. **50**, 3–14 (2014). https://doi.org/10.1016/j.patrec.2014.03.023

[11] Schaper, M.-M., Santos, M., Malinverni, L., Berro, J. Z. & Pares, N. Learning about the past through situatedness, embodied exploration and digital augmentation of cultural heritage sites. Int. J. Hum.-Comput. Stud. **114**, 36–50 (2018). https://doi.org/10.1016/j.ijhcs.2018.01.003

[12] Machidon, O. M., Duguleana, M. & Carrozzino, M. Virtual humans in cultural heritage ICT applications: A review. J. Cult. Herit. **33**, 249–260 (2018). https://doi.org/10.1016/j.culher.2018.01.007

[13] Leopardi, A. et al. X-reality technologies for museums: a comparative evaluation based on presence and visitors experience through user studies. J. Cult. Herit. **47**, 188–198 (2021). https://doi.org/10.1016/j.culher.2020.10.005

[14] Zhou, Y., Chen, J. & Wang, M. A meta-analytic review on incorporating virtual and augmented reality in museum learning. Educ. Res. Rev. **36**, 100454 (2022). https://doi.org/10.1016/j.edurev.2022.100454

[15] Innocente, C., Ulrich, L., Moos, S. & Vezzetti, E. A framework study on the use of immersive XR technologies in the cultural heritage domain. J. Cult. Herit. **62**, 268–283 (2023). https://doi.org/10.1016/j.culher.2023.06.001

[16] Chen, W., Li, T. & Zhang, Y. Embodied cognition model for museum gamification cultural heritage communication a grounded theory study. npj Herit. Sci. **13**, 239 (2025). https://doi.org/10.1038/s40494-025-01821-9

[17] Wu, Y. et al. KaiBiLi: gesture-based immersive virtual reality ceremony for traditional Chinese cultural activities. Vis. Comput. Ind. Biomed. Art 8, 24 (2025). https://doi.org/10.1186/s42492-025-00205-x

[18] Wang, X., Fu, J., Luo, S., Zhao, K. & Duan, Q. XR-based cultural heritage information spaces: Towards a new paradigm of information encoding and decoding guided by embodied cognition. Inf. Process. Manag. **63**, 104445 (2026). https://doi.org/10.1016/j.ipm.2025.104445

[19] Zhang, Y. et al. CECL: a chain-based framework for learning-oriented extended reality in cultural heritage and its empirical evaluation. npj Herit. Sci.; https://doi.org/10.1038/s40494-026-02726-x (2026).

[20] Yu, Z., Zhou, X., Yu, Z., Park, J. H. & Ma, J. iMuseum: A scalable context-aware intelligent museum system. Comput. Commun. **31**, 4376–4382 (2008). https://doi.org/10.1016/j.comcom.2008.05.004

[21] Ghiani, G., Paternò, F., Santoro, C. & Spano, L. D. UbiCicero: A location-aware, multi-device museum guide. Interact. Comput. **21**, 288–303 (2009). https://doi.org/10.1016/j.intcom.2009.06.001

[22] Ruotsalo, T. et al. SMARTMUSEUM: A mobile recommender system for the Web of Data. J. Web Semant. **20**, 50–67 (2013). https://doi.org/10.1016/j.websem.2013.03.001

[23] Hwang, G.-J., Yang, T.-C., Tsai, C.-C. & Yang, S. J. H. A context-aware ubiquitous learning environment for conducting complex science experiments. Comput. Educ. **53**, 402–413 (2009). https://doi.org/10.1016/j.compedu.2009.02.016

[24] Bektaş, K., Strecker, J., Mayer, S. & Garcia, K. Gaze-enabled activity recognition for augmented reality feedback. Comput. Graph. **119**, 103909 (2024). https://doi.org/10.1016/j.cag.2024.103909

[25] Chen, C.-Z., Wang, W.-S. & Huang, Y.-M. Designing adaptive feedback for hands-on learning in VR: Feedback intervention theory. Educ. Inf. Technol. **31**, 4115–4139 (2026). https://doi.org/10.1007/s10639-026-13968-4

[26] Huang, Y. Y., Yu, S. S., Chu, J. J., Fan, H. H. & Du, B. B. Using knowledge graphs and deep learning

algorithms to enhance digital cultural heritage management. Herit. Sci. **11**, 204 (2023). https://doi.org/10.1186/s40494-023-01042-y

[27] Lu, L. & Li, M. Development of a virtual interactive system for Dahua Lou loom based on knowledge ontology-driven technology. Herit. Sci. **11**, 178 (2023). https://doi.org/10.1186/s40494-023-01027-x

[28] Xu, L. et al. Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. Herit. Sci. **12**, 118 (2024). https://doi.org/10.1186/s40494-024-01231-3

[29] Yuan, H., Li, Y., Wang, B., Liu, K. & Zhang, J. Knowledge graph-based intelligent question answering system for ancient Chinese costume heritage. npj Herit. Sci. 13, 198 (2025). https://doi.org/10.1038/s40494-025-01776-x

[30] Wan, J., Li, X., Zhang, H., Zou, A. & Wang, R. An LLM-based QA system for Chinese Painting and Calligraphy with Knowledge Graphs and external documents. npj Herit. Sci. **13**, 643 (2025). https://doi.org/10.1038/s40494-025-02219-3

[31] Chen, Y. & Zhang, S. A knowledge-graph-guided interactive system with controllable generation for enhancing the interpretation of Song Dynasty ceramics. npj Herit. Sci.; https://doi.org/10.1038/s40494-026-02587-4 (2026).

[32] Hu, Y. MuseRAG++: a deep retrieval-augmented generation framework for semantic interaction and multi-modal reasoning in virtual museums. Sci. Rep. **16**, 25218 (2026). https://doi.org/10.1038/s41598-026-55700-9

[33] Fan, Y. et al. Explainable semantic question answering system for cultural heritage: a framework based on a dual-layer knowledge graph and self-RAG-inspired adaptive retrieval—a case study of Liao Dynasty tomb mural costume. Sci. Rep.; https://doi.org/10.1038/s41598-026-63351-z (2026).

[34] Wang, Y., Xu, P., Zhao, H. & Xu, K. Modeling dual temporal dependencies for temporal knowledge graph reasoning on Chinese cultural heritage. npj Herit. Sci. **14**, 463 (2026). https://doi.org/10.1038/s40494-026-02586-5

[35] Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M.-W. Retrieval Augmented Language Model Pre-Training. Proc. Mach. Learn. Res. 119, 3929–3938 (2020).

[36] Karpukhin, V. et al. Dense passage retrieval for open-domain question answering. In Proc. EMNLP 2020 6769–6781 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.emnlp-main.550

[37] Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst. **33**, 9459–9474 (2020). https://doi.org/10.48550/arXiv.2005.11401

[38] Edge, D. et al. From local to global: a graph RAG approach to query-focused summarization. Preprint at https://doi.org/10.48550/arXiv.2404.16130 (2024).

[39] Peng, B. et al. Graph Retrieval-Augmented Generation: A Survey. ACM Trans. Inf. Syst. **44**, 35 (2025). https://doi.org/10.1145/3777378

[40] Singh, A., Ehtesham, A., Kumar, S., Khoei, T. T. & Vasilakos, A. V. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. Preprint at https://doi.org/10.48550/arXiv.2501.09136 (2025).

[41] Needham, J., Ho, P. Y., Lu, G.-D. & Sivin, N. Science and Civilisation in China, Vol. 5, Part 2: Spagyrical Discovery and Invention: Magisteries of Gold and Immortality (Cambridge Univ. Press, 1974).

[42] Sivin, N. Chinese Alchemy: Preliminary Studies (Harvard Univ. Press, 1968).

[43] Pregadio, F. Great Clarity: Daoism and Alchemy in Early Medieval China (Stanford Univ. Press, 2006). https://doi.org/10.11126/stanford/9780804751773.001.0001

[44] China Chemical Industry Museum. Zhongguo Huagong Tongshi: Gudai Juan [A General History of Chinese Chemical Industry: Ancient Volume] (Chemical Industry Press, 2022).

[45] Ho, P.-Y. & Needham, J. The Laboratory Equipment of the Early Mediaeval Chinese Alchemists. Ambix 7, 58–112 (1959). https://doi.org/10.1179/amb.1959.7.2.58

[46] Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with GPUs. IEEE Trans. Big Data 7, 535–547 (2021). https://doi.org/10.1109/TBDATA.2019.2921572

[47] Chen, J. et al. BGE M3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. Preprint at https://doi.org/10.48550/arXiv.2402.03216 (2024).

[48] Hart, S. G. & Staveland, L. E. Development of NASA-TLX: results of empirical and theoretical research. Adv. Psychol. 52, 139–183 (1988). https://doi.org/10.1016/S0166-4115(08)62386-9

[49] Laugwitz, B., Held, T. & Schrepp, M. Construction and evaluation of a user experience questionnaire. In HCI and Usability for Education and Work 63–76 (Springer, 2008). https://doi.org/10.1007/978-3-540-89350-9_6

[50] Cohen, J. Statistical Power Analysis for the Behavioral Sciences (Lawrence Erlbaum Associates, 1988).