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.

ArchaeoBench: a large-scale archaeological question-answering benchmark for large language models

Xiaoquan Xie
Hangzhou City University
Linming Li
Hangzhou City University
Wei Ren
Hangzhou City University
Wei Zhang
Hangzhou City University
Fan Jiang
Hangzhou City University
Miaoyan Chen
Hangzhou City University
Bing Qi
Shanxi University
Kui Su
Hangzhou City University
Jun Wang
wangjun@hzcu.edu.cn
Hangzhou City University
Zhichun Jing
University of British Columbia
Dalun Gao
Shanxi University

Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11624722/v1

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

Additional Declarations: No competing interests reported.

ArchaeoBench: a large-scale archaeological question-answering benchmark for large language models

Xiaoquan Xie[1], Linming Li[2], Wei Ren[3], Wei Zhang[1], Fan Jiang[3], Miaoyan Chen[2], Bing Qi[4], Kui Su[1, 2]✉, Jun Wang[1]✉, Zhichun Jing[5]✉ & Dalun Gao[4]✉

Abstract: Despite the rapid adoption of large language models (LLMs) in data-intensive disciplines, archaeology lacks public, large-scale benchmarks, and its spatio-temporally fragmented knowledge resists integration into model parameters. Here we introduce ArchaeoBench, a hierarchically structured benchmark built from 63 official excavation reports across 20 Chinese provinces, comprising over 9.2 million training question-answer pairs and 1,698 evaluation questions spanning multiple cognitive levels. Fine-tuning seven open-source LLMs on this benchmark yields substantial gains, approaching commercial baselines on several question types, whereas chain-of-thought prompting brings only marginal improvements on fact- and text-grounded questions, suggesting that archaeological question answering is constrained primarily by access to reliable domain knowledge rather than reasoning depth. Motivated by this finding, we propose a knowledge-graph-based retrieval-augmented generation framework that injects curated structured evidence to enhance spatio-temporal and cultural-linkage reasoning. ArchaeoBench converts excavation reports into reusable, machine-readable knowledge infrastructure for the discipline.

---

[1]School of Computer and Computing Science, Hangzhou City University, Hangzhou, China. [2]Supercomputing Center of Hangzhou City University, Hangzhou, China. [3]School of Art and Archaeology, Hangzhou City University, Hangzhou, China. [4]School of Archaeology and Museology, Shanxi University, Taiyuan, China. [5]Canada Research Chair in Pacific Asia Archaeology Department of Anthropology University of British Columbia 6303 NW Marine Drive Vancouver, BC V6T1Z1, Canada. ✉e-mail: suk@hzcu.edu.cn; wangjun@hzcu.edu.cn; jingzh@mail.ubc.ca; 583498488@qq.com; qibing@sxu.edu.cn.

Introduction

Archaeology is a data-intensive discipline that relies on massive volumes of unstructured data, including texts, images, and spatial records drawn from heterogeneous sources across different subfields. These materials are not only dispersed but also lack standardized formats and consistent scale[1,2]. Among them, excavation reports are arguably the most critical textual resources. Produced after each field excavation, they document site discoveries in detail and frame the academic interpretations that guide subsequent research. Yet the knowledge contained within them remains largely locked inside the reports themselves. It is fragmented across excavations, institutions, and publication traditions, resisting the systematic aggregation and cross-site synthesis that other text-rich disciplines have achieved. The sheer volume of these reports makes manual review and synthesis prohibitively time-consuming.

The recent advancement of Large Language Models (LLMs) has opened a new path. With enhanced natural language capabilities, LLMs can now parse and reason over lengthy technical documents, making automated processing of archaeological texts a realistic prospect. These capabilities have already driven successful applications in fields such as biology[3] , medicine[4] , law[5] , and astronomy[6] , as well as multimodal AI applications in cultural heritage and archeology[7] . A common strategy across these efforts involves constructing high-quality domain-specific datasets and fine-tuning general-purpose models to adapt them to specialized tasks. Whether this paradigm transfers to archaeology, however, remains an open empirical question, as archaeology differs from biomedicine and law in two fundamental respects.

The first difference concerns data. Biomedicine and law have long maintained curated corpora and community benchmarks, whereas archaeology lacks public, large-scale, and high-quality datasets for model training and evaluation. Its textual resources remain largely unstructured and inaccessible. The second difference concerns the nature of archaeological knowledge itself. Such knowledge is deeply fragmented across excavations, institutions, and publication traditions, carrying strong spatio-temporal dependencies alongside considerable semantic ambiguity. Information drawn from different strata, periods, and cultural contexts resists straightforward integration into the parameter space

of LLMs through conventional fine-tuning. These two challenges, namely data scarcity and knowledge complexity, jointly bottleneck the adaptation of LLMs to archaeology and motivate solutions that go beyond generic fine-tuning strategies.

To address these challenges, we constructed ArchaeoBench, a large-scale archaeological dataset built upon 63 official excavation reports from 20 provincial-level administrative regions across China. The reports were deliberately sampled to span the major subfields of the discipline, such as scientific archaeology, lithic analysis, settlement archaeology, mortuary archaeology, ceramic studies, jade studies, bronze studies, and architectural archaeology. The dataset comprises two components with strict provenance separation: a training corpus of over 9,000,000 Question-Answer (QA) pairs, and an evaluation benchmark of 1,698 questions in five formats, namely Choice Questions (CQs), True/False questions (TFs), Filling in Blanks (FBs), Terminologies (TMs), and Short Answers (SAs), which are designed to assess model capabilities across cognitive levels from factual recall and comprehension to synthesis and critical analysis. All evaluation questions were constructed under a decontamination protocol that prevents near-duplicate content from appearing in both the training corpus and the benchmark. Although the present dataset draws on Chinese excavation reports, the construction protocol is domain-driven rather than language-driven, allowing direct transfer to excavation report traditions in other regions.

Based on the ArchaeoBench dataset, we fine-tuned and systematically evaluated seven open-source LLMs, including MiniCPM4-8B[8] , MiniCPM5-2B[9] , Qwen3-8B[10] , Meta-Llama-3-8B[11] , Mistral-7B[12] , ChatGLM3-6B[13] , and Nanbeige4.2-3B[14] . We additionally evaluated three state-of-the-art (SOTA) commercial LLMs, namely DeepSeek-V4-Pro[15] , Qwen3.7-Max[16] and GLM-5.2[17] , as strong baselines without fine-tuning. Beyond benchmarking, we treat this ten-model comparison as an instrument for probing how current models encode and retrieve archaeological knowledge. Beyond performance along, the results speak to the fundamental nature of archaeological question answering. Finally, using the Nanlou Site[18] as a representative case study, we preliminarily explore whether a site-specific knowledge graph can be coupled with Retrieval-Augmented Generation (RAG)[19] so as to externalize the spatio-temporal and cultural-relational structure that fine-

tuning alone cannot capture, thereby further enhancing model performance on archaeological tasks.

Extensive experiments yield three principal findings. First, fine-tuned open-source models achieve substantial performance gains over their non-fine-tuned baselines, with several becoming competitive with or even surpassing commercial models. This demonstrates that domain knowledge injection through targeted fine-tuning is an effective route for archaeology, even at small or modest model scales. Second, for questions dominated by factual recall and text-grounded comprehension, which constitute the majority of routine archaeological QA, the benefits of Chain-of-Thought (CoT) reasoning are marginal. This observation suggests that much of archaeological question answering functions primarily as retrieval and synthesis over domain corpora rather than multi-step reasoning. It carries a direct implication for the design of digital-heritage systems: investing in better knowledge organization and retrieval infrastructure may matter more than investing in reasoning capabilities. Third, our preliminary evaluation shows that coupling a site-specific knowledge graph with a RAG pipeline further enhances model performance, particularly on tasks that require spatio-temporal reasoning and cultural linkage analysis. Knowledge graphs therefore offer a promising route to mitigating the knowledge fragmentation that fine-tuning alone cannot resolve.

Taken together, this work makes three contributions to computing for cultural heritage. First, ArchaeoBench is, to our knowledge, the first large-scale archaeological benchmark dataset featuring hierarchical QA pairs, multi-format evaluation questions, and explicit train-evaluation decontamination. Second, we provide a systematic empirical analysis spanning ten models, which identifies both the capabilities and the current limitations of LLMs in archaeology and characterizes the retrieval-dominant nature of archaeological QA. Third, we propose a knowledge-graph-augmented RAG framework, validated on a targeted case study, that offers a concrete direction for integrating fragmented spatio-temporal knowledge into archaeological LLM applications.

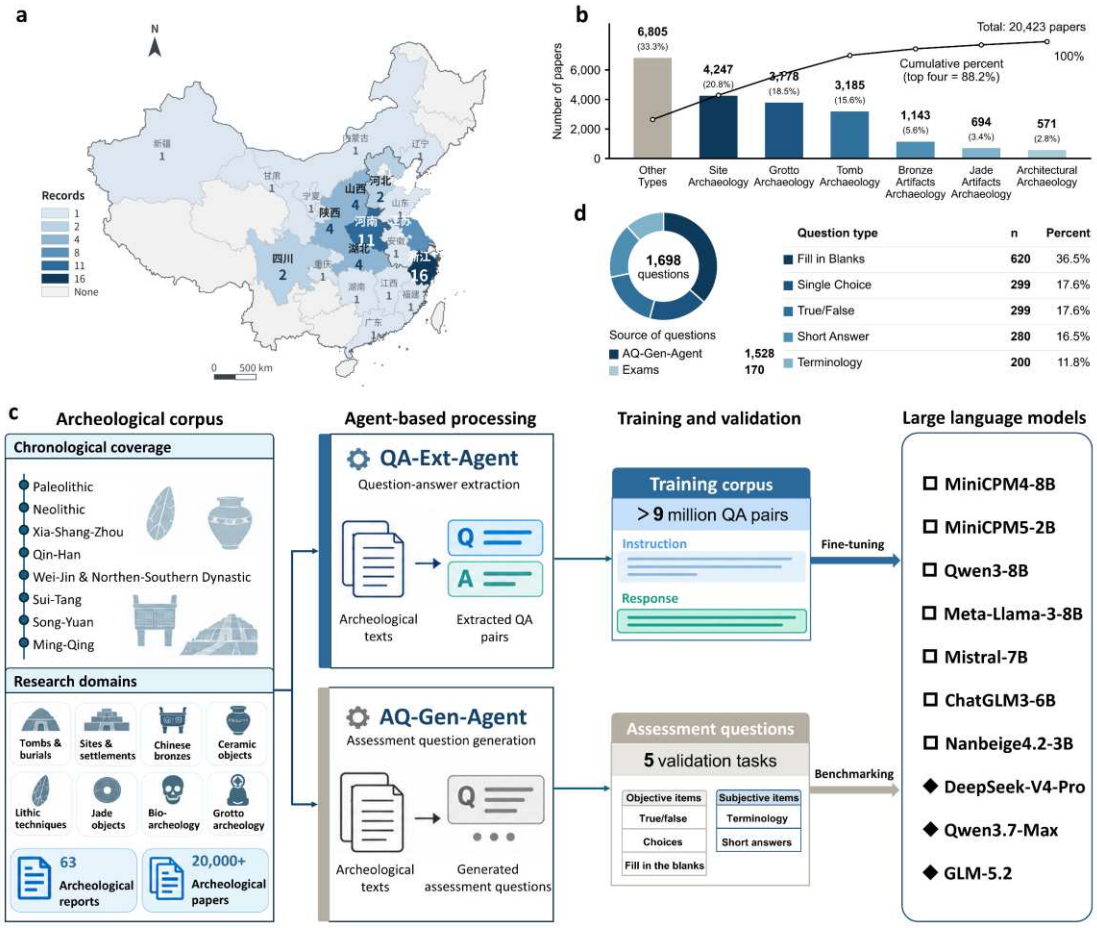


Fig. 1 Study overview. Data collection from two primary corpora: **a** Archaeological reports from 63 official excavations across 20 provincial-level administrative regions of China and **b** over 20,000 archaeological research papers covering diverse research subfields. **c** The overall study pipeline comprises two main stages: 1) Agent-based processing for automated QA pair extraction, where two specialized agents, the QA-Ext-Agent and the AQ-Gen-Agent, collaboratively parse the raw corpus to generate structured QA pairs for training and multi-format assessment questions for evaluation and 2) Model adaptation and evaluation, where multiple representative SOTA LLMs are fine-tuned on the generated QA pairs and benchmarked against the assessment questions using standardized objective metrics. **d** The assessment question distribution among five tasks.

Results

Study overview

Fig. 1 illustrates the overall framework of the ArchaeoBench. In the corpus construction stage, we assembled textual resources from 63 official excavation reports across 20 provincial-level administrative regions of China (Fig. 1a) and over 20,000 archaeological research papers, covering diverse subfields such as Tomb, Sites, and Grotto (Fig. 1b). These archaeological corpora are characterized not only by their extensive geographical coverage and broad disciplinary scope, but also by a vast historical span,

ranging from the Stone Age all the way to the Yuan, Ming, and Qing dynasties.

Building on this extensive and diverse textual foundation, our study pipeline comprises two main stages (Fig. 1c):1) Agent-based processing for automatic QA pair extraction, and 2) model training and validation. In the first stage, we designed two specialized agents, the QA Pair Extraction Agent (QA-Ext-Agent) and the Assessment Question Generation Agent (AQ-Gen-Agent), which automatically parse the raw corpus to produce 9,241,289 structured QA pairs for training (train set) and 1,698 multi-format assessment questions for evaluation (test set, Fig. 1d), respectively. The test set comprises 90% questions generated by the AQ-Gen-Agent and 10% sourced from real-world archaeology entrance exams across multiple universities (Fig. 1d). Notably, to ensure the factual accuracy and scholarly quality of the generated content, three senior archaeology experts were involved in a multi-round review process to verify the QA pairs and assessment questions. Table 1 summarizes the statistical information of the QA pairs. Further details on dataset construction and statistics are provided in the Methods section.

Table 1. Statistical information of the archaeological QA pairs

Items	Value
Average question length	38.7 tokens
Average solution length	213.6 tokens
Clauses per example	2.7 clauses
Train set size	9,241,289 QA pairs (2.67 billion tokens)
Test set size	1,698 questions (0.34 million tokens)

Supplementary Fig. 1 presents representative QA pair examples from the training dataset, using the Jingnan Temple Site[20] and the Tomb of the Lady of the Qi State[21] as illustrative cases. Each question is formulated with complete contextual information, including location, time period, and the specific subject in question, to minimize semantic ambiguity. Supplementary Fig. 2 illustrates the five question formats included in the evaluation benchmark, where each question is annotated with both the ground-truth

answer and its corresponding archaeological subfield.

Finally, in the model training and evaluation stage, we fine-tuned multiple representative SOTA LLMs on the generated QA pairs and benchmarked their performance on the 1,698 assessment questions using standardized evaluation metrics.

Model selection and fine-tuning

To comprehensively evaluate the value of our dataset and the capabilities of existing general-purpose LLMs in archaeological question answering and reasoning, we selected multiple open-source and commercial models for comparative validation (see Table 2).

Table 2. Large language models selected for evaluation

Model	Params	Window size	Open source	Fine-tuning
MiniCPM4-8B	8.19B	32K	Yes	LoRA vs DoRA
MiniCPM5-2B	2.52B	128K	Yes	LoRA vs DoRA
Qwen3-8B	8.2B	32K	Yes	LoRA vs DoRA
Meta-Llama-3-8B	8.03B	8k	Yes	LoRA vs DoRA
Mistral-7B	7.25B	32K	Yes	LoRA vs DoRA
ChatGLM3-6B	6.24B	32k	Yes	LoRA vs DoRA
Nanbeige4.2-3B	4.17B	256k	Yes	LoRA vs DoRA
DeepSeek-V4-Pro	49B	1M	No	No fine-tuning
Qwen-3.7-Max	45B	1M	No	No fine-tuning
GLM-5.2	40B	1M	No	No fine-tuning

For the open-source LLMs, we compared two fine-tuning strategies: Low-Rank Adaptation (LoRA)[22] , which approximates weight updates using low-rank decomposition, and Decomposed Low-Rank Adaptation (DoRA)[23] , which further separates the pre-trained weight matrix into magnitude and directional components and applies low-rank adaptation solely to the latter. We chose these two methods because they represent the most widely adopted Parameter-Efficient Fine-Tuning (PEFT) approach.

For the commercial LLMs, we evaluated their performance under two configurations:

with CoT and without CoT (i.e., No-CoT) reasoning, to comprehensively assess the capabilities of these SOTA commercial LLMs on archaeological tasks. Notably, to ensure a fair comparison, we conducted offline evaluations to prevent the models from accessing the internet.

Performance gains on objective questions

For the objective questions (i.e., CQs, TFs, and FBs), we employ *accuracy* as an intuitive metric to evaluate model performance, which is defined as follows:

$$Accuracy = \frac{N_{\text{correct}}}{N_{\text{obj}}} \quad (1)$$

where N_{correct} denotes the number of correctly answered questions, and N_{obj} represents the total number of objective questions.

The performance of all open-source LLMs is presented in Table 3. From the results, the following two main conclusions can be drawn: 1) There is a significant performance disparity among these models with varying parameter sizes. For instance, the Nanbeige4.2-3B model achieved only 38.80% accuracy on TFs, whereas the Qwen3-8B model reached a remarkable 96.32%; 2) Both LoRA and DoRA fine-tuning generally improved the accuracy of most baseline models across the three question types, though the gains varied considerably. Specifically, LoRA fine-tuning yielded an average improvement of 5.73% on CQs, while DoRA fine-tuning resulted in an average gain of 5.07%. The improvements were more pronounced on TFs, with LoRA and DoRA fine-tuning achieving average increases of 19.64% and 18.59%, respectively.

Table 3. Performance comparison of LoRA and DoRA fine-tuning across open-source LLMs on objective evaluation questions （Accuracy%）

Model	Choice question	True/False	Filling in blanks
MiniCPM4-8B	87.63%	81.61%	37.26%
MiniCPM4-8B-LoRA	90.30% (+2.67%)	87.96% (+6.35%)	39.35% (+2.09%)
MiniCPM4-8B-DoRA	88.96% (+1.33%)	88.96% (+7.35%)	40.16% (+2.90%)
MiniCPM5-2B	80.27%	95.99%	33.55%
MiniCPM5-2B-LoRA	80.60% (+0.33%)	96.32% (+0.33%)	36.45% (+2.90%)

MiniCPM5-2B-DoRA	80.94% (+0.67%)	96.32% (+0.33%)	35.00% (+1.45%)
Qwen3-8B	88.29%	96.32%	43.13%
Qwen3-8B-LoRA	92.98% (+4.69%)	96.32% (+0.00%)	48.63% (+5.50%)
Qwen3-8B-DoRA	92.98% (+4.69%)	96.32% (+0.00%)	45.88% (+2.75%)
Meta-Llama-3-8B	39.46%	47.83%	43.23%
Meta-Llama-3-8B-LoRA	41.47% (+2.01%)	84.28% (+36.45%)	53.71% (+10.48%)
Meta-Llama-3-8B-DoRA	40.47% (+1.01%)	74.92% (+27.09%)	54.52% (+11.29%)
Mistral-7B	71.57%	58.19%	27.30%
Mistral-7B-LoRA	89.63% (+18.06%)	95.65% (+37.46%)	46.69% (+19.39%)
Mistral-7B-DoRA	87.96% (+16.39%)	94.98% (+36.79%)	48.47% (+21.17%)
ChatGLM3-6B	75.92%	91.78%	38.55%
ChatGLM3-6B-LoRA	80.60% (+4.68%)	93.84% (+2.06%)	48.55% (+10.00%)
ChatGLM3-6B-DoRA	80.60% (+4.68%)	94.52% (+2.74%)	47.42% (+8.87%)
Nanbeige4.2-3B	79.26%	38.80%	30.32%
Nanbeige4.2-3B-LoRA	86.96% (+7.70%)	93.65% (+54.85%)	30.97% (+0.65%)
Nanbeige4.2-3B-DoRA	85.95% (+6.69%)	94.65% (+55.85%)	29.84% (-0.48%)

Improvements on open-ended questions

We employ six widely adopted metrics to evaluate the quality of responses to the Terminology and Short Answers: $BLEU$[24], $ROUGE1$, $ROUGE2$, $ROUGEL$[25], $BERTScore$[26], and $CHIE_{\text{overall}}$[27]. While the first five metrics assess lexical overlap and semantic similarity, $CHIE_{\text{overall}}$ evaluates the comprehensive application quality of the model's output. It is calculated as the weighted average of four sub-metrics:

$$CHIE_{\text{overall}} = 0.35 \times CHIE_{\text{pre}} + 0.25 \times CHIE_{\text{Help}} + 0.25 \times CHIE_{\text{rel}} + 0.15 \times CHIE_{\text{con}}. \quad (2)$$

where $CHIE_{\text{pre}}$, $CHIE_{\text{Help}}$, $CHIE_{\text{rel}}$, and $CHIE_{\text{con}}$ measures the factual correctness, practical utility, alignment with the original query, and freedom from verbosity, respectively. Then, we computed the following weighted metric based on the above six indicators for evaluation of each response:

$$Score_{\beta} = 0.05 \times BLEU + 0.1 \times ROUGE1 + 0.05 \times ROUGE2 + 0.05 \times ROUGEL +$$

$$0.25 \times BERTScore + 0.5 \times CHIE_{\text{overall}}. \quad (3)$$

Table 4. Performance comparison of LoRA and DoRA fine-tuning across open-source LLMs on Terminology and Short answers (sum of $Score_{\beta}$)

Model	Terminology	Short Answers
MiniCPM4-8B	76.21	90.29
MiniCPM4-8B-LoRA	74.81 (-1.40)	93.87 (+3.58)
MiniCPM4-8B-DoRA	74.19 (-2.02)	94.20 (+3.91)
MiniCPM5-2B	73.93	88.84
MiniCPM5-2B-LoRA	75.20 (+1.27)	90.33 (+1.49)
MiniCPM5-2B-DoRA	74.80 (+0.87)	91.20 (+2.36)
Qwen3-8B	72.85	89.65
Qwen3-8B-LoRA	82.72 (+9.87)	94.12 (+4.47)
Qwen3-8B-DoRA	84.89 (+12.04)	94.84 (+5.19)
Meta-Llama-3-8B	58.89	76.64
Meta-Llama-3-8B-LoRA	62.11 (+3.22)	91.12 (+14.48)
Meta-Llama-3-8B-DoRA	62.36 (+3.47)	91.14 (+14.50)
Mistral-7B	64.77	86.08
Mistral-7B-LoRA	81.81 (+17.04)	101.06 (+14.98)
Mistral-7B-DoRA	87.64 (+22.87)	102.43 (+16.35)
ChatGLM3-6B	68.28	97.01
ChatGLM3-6B-LoRA	67.12 (-1.16)	100.28 (+3.27)
ChatGLM3-6B-DoRA	67.26 (-1.02)	100.12 (+3.11)
Nanbeige4.2-3B	53.92	72.88
Nanbeige4.2-3B-LoRA	67.87 (+13.95)	112.86 (+39.98)
Nanbeige4.2-3B-DoRA	61.91 (+7.99)	112.41 (+39.53)

Table 4 presents the sum of $Score_{\beta}$ for all model responses, calculated as $\sum_{i}^{N_{\text{open}}} Score_{\beta}^{i}$, where N_{open} denotes the total number of Terminology or Short Answer

questions. The results show that fine-tuning does not uniformly improve performance across all models. While most models benefit from LoRA and DoRA, several exhibit slight decreases (e.g., the performance of MiniCPM4-8B and ChatGLM3-6B on Terminology). On average, the performance gain on Short Answer questions is larger than that on Terminology questions. This may be attributed to the fact that Short Answer tasks require broader reasoning and synthesis capabilities, which can benefit substantially from instance-level fine-tuning, whereas Terminology tasks rely on precise rote memorization of domain definitions and are therefore less sensitive to such fine-tuning. More detailed metric comparisons for Qwen3-8B and Mistral-7B are presented as Radar charts in Fig. 2, where the average value of each metric is calculated across all responses. Notably, the metrics also show more pronounced improvements for Short Answer questions than for Terminology questions, further confirming the above conclusion.

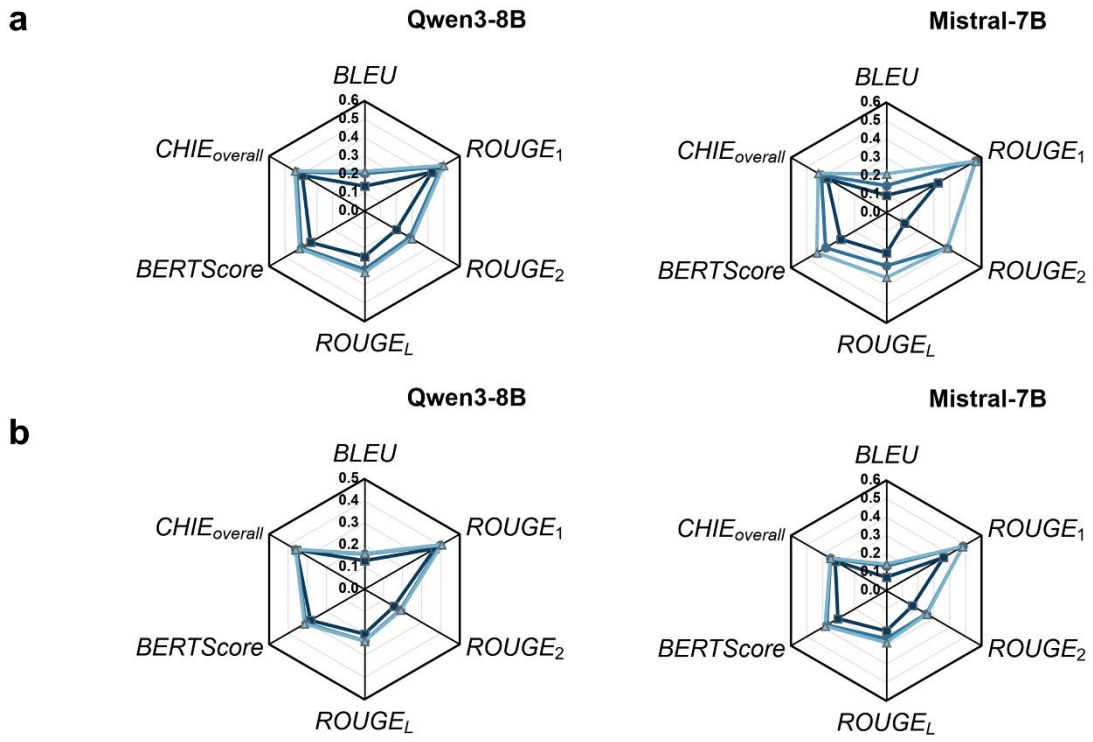


Fig. 2 Performance comparison of LLMs under baseline, LoRA, and DoRA fine-tuning strategies. **a** Performance on Terminology and **b** on Short Answers. Fine-tuning yields significantly greater improvements on Short Answers than on Terminology tasks. This discrepancy may be attributed to the fact that Short Answers demand broader reasoning and synthesis capabilities, which can benefit substantially from instance-level tuning, whereas Terminology questions rely on precise rote memorization and are therefore less sensitive to such fine-tuning.

Limited gains of Chain-of-Thought in SOTA Commercial LLMs

The aforementioned validations were conducted on relatively smaller open-source models, aiming to demonstrate that fine-tuning with our ArchaeoBench dataset significantly enhances archaeological QA capabilities. To further evaluate the performance of commercial LLMs and the impact of CoT prompting, we assessed three SOTA commercial models, i.e., DeepSeek-V4-Pro (49B), Qwen3.7-Max (45B), and GLM-5.2 (40B), under two distinct settings: No-CoT and with CoT.

Fig. 3 presents a performance comparison of these models across multiple tasks under both No-CoT and CoT settings. Two key conclusions can be drawn: 1) The performance of these commercial LLMs is significantly superior to that of the open-source models discussed earlier, likely attributable to their larger parameter sizes and exposure to more extensive training corpora; 2) The performance gap between the CoT and No-CoT modes is marginal, suggesting that CoT prompting offers limited advantages in the domain of archaeological question answering, as the tasks primarily involve factual knowledge retrieval rather than complex multi-step reasoning.

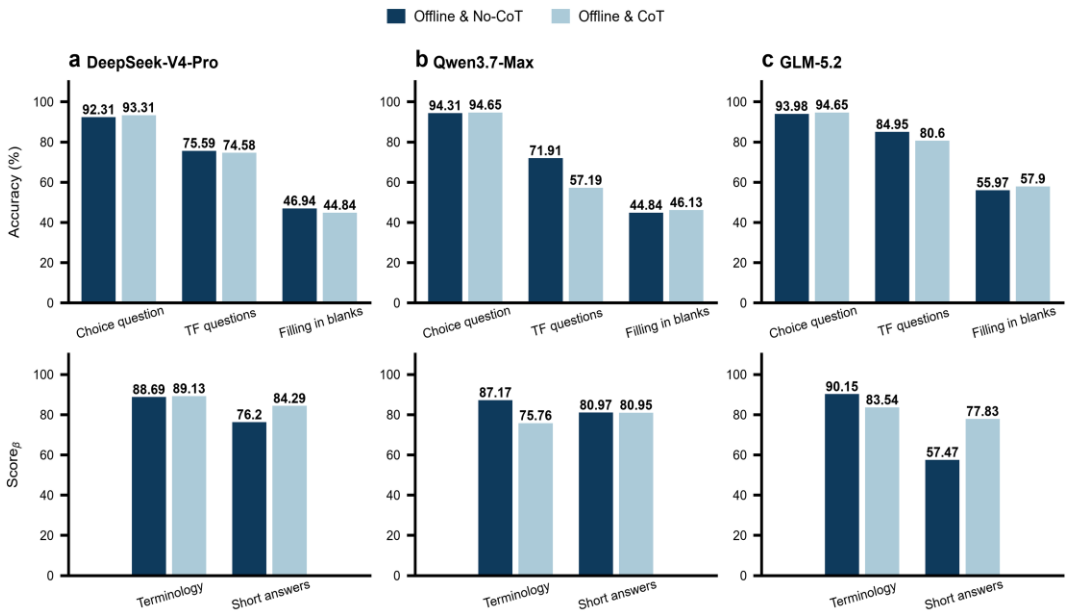


Fig. 3 Performance of SOTA commercial models under CoT vs No-CoT settings. **a** Results for DeepSeek-V4-Pro, where the top and bottom rows display the performance on objective questions (evaluated by *accuracy*) and subjective questions (evaluated by $Score_\beta$), respectively; **b** and **c** show the corresponding results for Qwen3.7-Max and GLM-5.2, following the same layout as **a**. The results indicate that CoT prompting does not yield significant improvements on our benchmark.

Examples of model response

To provide an intuitive demonstration of the fine-tuning effects, Fig.4 presents the responses of the Mistral-7B base model, the LoRA fine-tuned version, and the DoRA fine-tuned version to a Terminology question. Additionally, Qwen3.7 was employed as an evaluator to assign scores (on a 0–100 scale) and provide qualitative assessments for each response, with the results displayed beneath the corresponding outputs. As can be observed, the fine-tuned models substantially outperform the base version, with the LoRA and DoRA variants achieving scores of 55 and 30, respectively, compared to only 10 for the base model. As noted in the evaluations, the base model's response contains fundamental factual errors, describing content entirely unrelated to the Xuetangliangzi Site; the entire response constitutes a model hallucination. In contrast, the LoRA and DoRA responses are factually accurate in terms of geographic location, chronological period, and key archaeological findings.

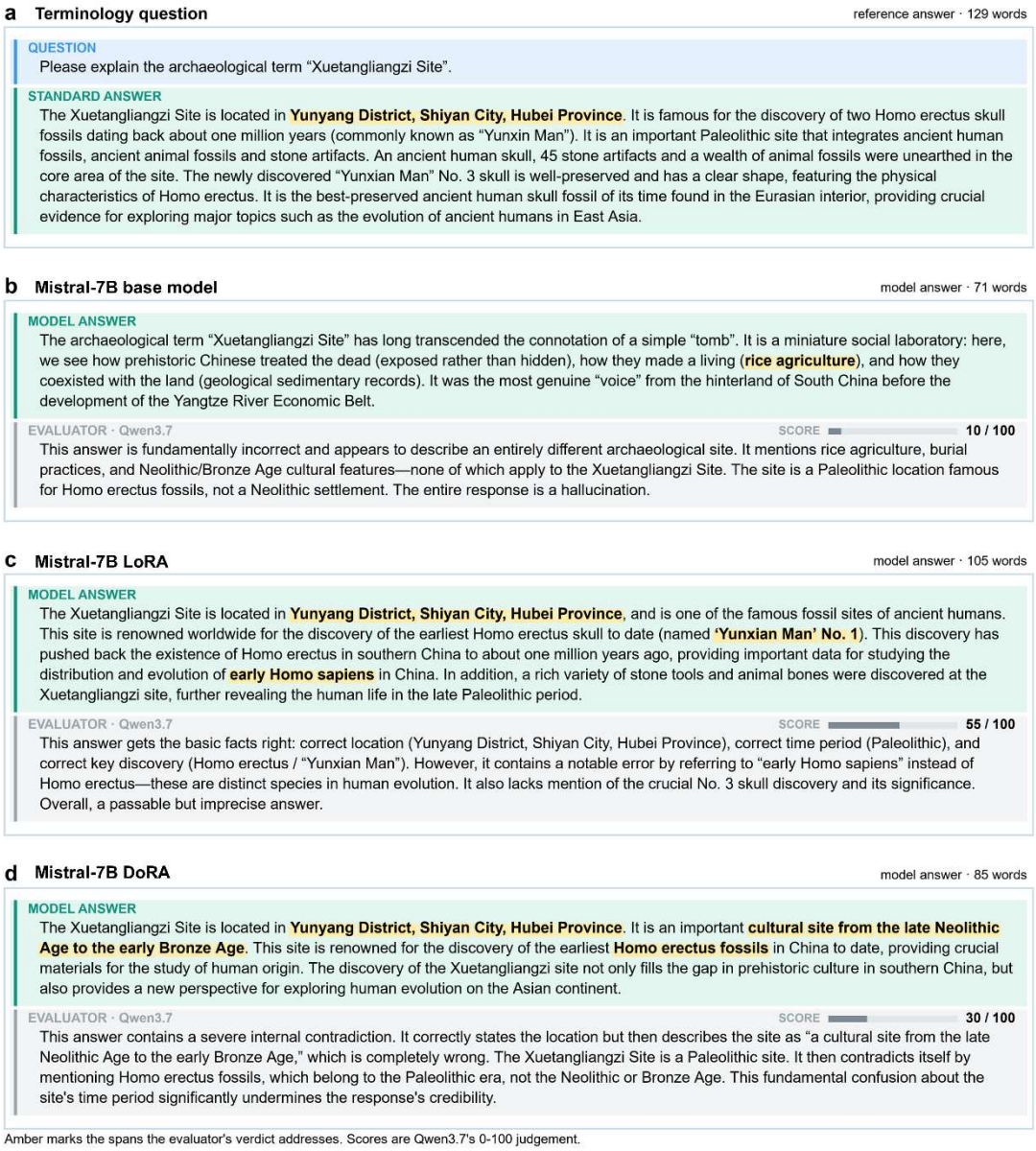


Fig. 4 Example responses to a Terminology question based on the Xuetangliangzi Site. a The question and ground-truth answer. **b** The base version of Mistral-7B without fine-tuning is entirely incorrect, whereas after fine-tuning with LoRA (**c**) and DoRA (**d**), it produces more accurate answers, correctly identifying the geographic location of the archaeological site. Evaluation by Qwen3.7 yields consistent results: the LoRA fine-tuned version achieves a score of 55, significantly outperforming the base version with a score of only 10.

Knowledge graph-based RAG can reduce hallucinations

While fine-tuning significantly improves the precision of model responses, hallucinations may still persist, particularly regarding factual knowledge such as detailed quantitative data about archaeological sites. To mitigate incorrect responses, we introduce knowledge graph-based RAG, validated through a representative case study, namely the

Nanlou Site. Fig. 5 illustrates the hierarchical knowledge graph of the Nanlou Site constructed in this study. Specifically, Fig. 5a presents site-level information, such as chronology, cultural affiliation, geographic location, and excavation year. Fig. 5b details the excavated tombs, and Fig. 5c displays information on excavated ash pits along with their associated unearthed artifacts.

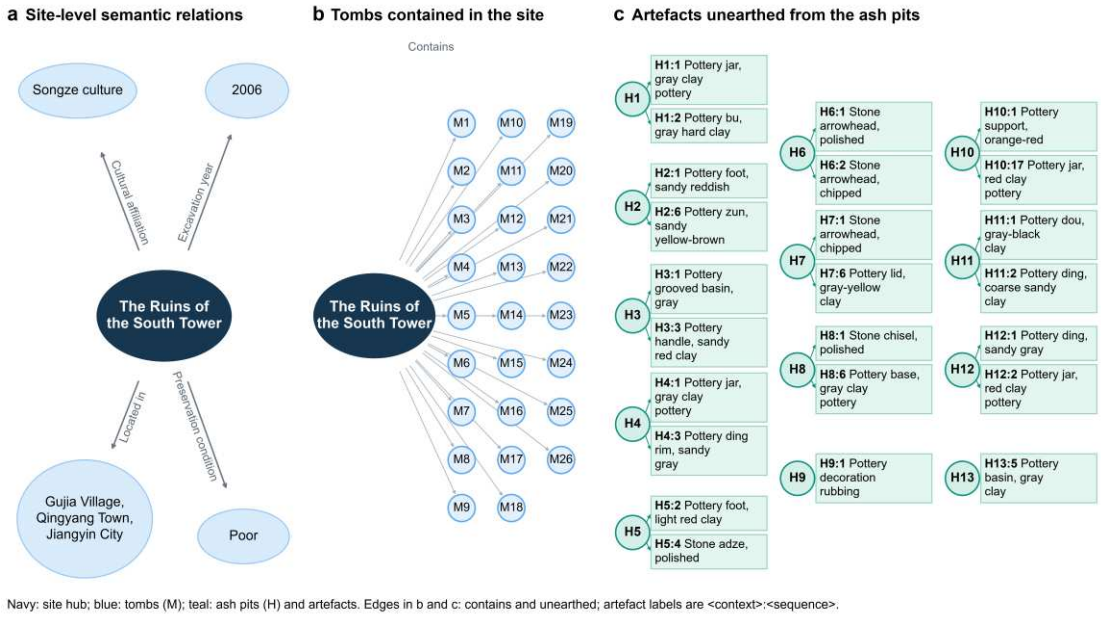


Fig. 5 Hierarchical knowledge graphs constructed for the Nanlou Site. **a** Overview of basic site-level attributes. **b** Inventory of associated tombs. **c** Excavation details of ash pits (only a subset is shown).

Our experimental results demonstrate that leveraging these knowledge graphs to construct RAG prompts significantly mitigates model hallucinations. As illustrated in Fig. 6b and c, in the absence of RAG, the fine-tuned model yields entirely erroneous responses, resulting in zero scores. However, upon incorporating the RAG mechanism, both fine-tuned models achieve accurate question answering. Notably, details concerning site locations and quantitative data, such as the site's area and dimensions highlighted in green, align perfectly with the ground truth, with scores reaching as high as 90.



Fig. 6 Knowledge Graph-based RAG reduces model hallucinations. **a** shows the question and the standard answer. **b** and **c** present the responses from Qwen3-8B with LoRA and DoRA fine-tuning, respectively. Without knowledge graph enhancement, both models exhibit hallucinations and provide incorrect answers. In contrast, **d** and **e** correspond to the fine-tuned versions with knowledge graph enhancement, which significantly mitigate hallucinations, yielding precise responses regarding site locations and quantitative data.

Discussion

This study presents ArchaeoBench, a large-scale archaeological QA dataset and evaluation benchmark for LLMs. Its primary contribution lies in transforming archaeological records into computable, reusable knowledge infrastructure. Unlike prior works limited to

specific tasks (e.g., archaeological named-entity recognition[28]), or to specialized regional and numerical records (e.g., the radiocarbon datasets of eastern Africa[29] and p3k14c[30]), official excavation reports constitute authoritative field archives yet remain locked within long-form narratives that are difficult for large-scale computational analysis. By decomposing these reports into hierarchically structured QA pairs and multi-format assessment questions, ArchaeoBench converts unstructured corpora into machine-readable knowledge while preserving source provenance. Its hierarchical design mirrors archaeological reasoning, progressing from site factual recall to stratigraphic, typological, and interpretive synthesis. The construction protocol, combining official-source sampling and train–evaluation decontamination, establishes a domain-driven quality standard transferable to excavation-report traditions across regions and languages. Beyond benchmarking, ArchaeoBench serves as a curated digital knowledge base supporting research synthesis, terminological standardisation and archaeological training.

Parameter-efficient fine-tuning methods, particularly LoRA and DoRA, have been widely adopted to adapt large language models to vertical domains such as healthcare[31] , law[32] , and finance[33] , achieving strong task-specific performance with minimal computational cost. We adopted LoRA and DoRA to adapt seven open-source LLMs to this archaeological corpus. Consistent performance gains across models confirm the dataset's training utility. Two practical implications emerge for heritage practice. First, a corpus of 2.67 billion tokens suffices to adapt small models (2B-8B parameters), meaning heritage institutions do not require frontier commercial models or heavy computational resources. Locally deployed domain-adapted open models are especially valuable for unpublished or restricted excavation reports that cannot be uploaded to external cloud services. Second, strong gains achieved by training fewer than 0.7% of model parameters (as little as 0.11% for the 8B models) indicate domain knowledge grounding, not raw model scale, is the main constraint for archaeological QA, lowering entry barriers for resource-limited research groups. Nevertheless, hallucination persists: fine-tuned models still produce factually incorrect outputs in fine-grained archaeological queries.

We further evaluated commercial LLMs including DeepSeek-V4-Pro, Qwen3.7-Max, and GLM-5.2. CoT prompting yield only marginal performance improvements. This aligns

with recent findings: CoT[34] benefits multi-step symbolic reasoning tasks such as mathematics, while archaeological QA is predominantly knowledge-retrieval intensive. The limited CoT gain demonstrates that the key bottleneck is not reasoning capacity, but reliable factual grounding within archaeological contexts. Methodologically, this implies diminishing returns from ever-larger reasoning models; greater gains come from investing in structured archives, standard terminology and controlled vocabularies for excavation records. It also warns against uncritical deployment of generic chatbots for archaeological reference work, where fluent yet ungrounded answers constitute measurable hallucination captured by this benchmark.

To mitigate hallucination in site-specific queries, we introduced a site-level knowledge graph to impose explicit structural constraints. The Nanlou Site case illustrates that fine-tuned models are prone to geographical and spatiotemporal hallucinations, whereas the knowledge graph[35] anchors generation to verified factual triples within a RAG pipeline. This encodes core archaeological relations: stratigraphic sequence, typological comparison, spatial adjacency and cultural affiliation. Represented as inspectable triples rather than implicit model parameters, the graph provides auditable knowledge for domain experts, enabling human-machine collaboration unavailable from fine-tuning alone. Integrating fine-tuned LLMs with site-specific knowledge graphs provides a scalable framework for constructing digital site archives. Once standardized and interoperable across different sites, such graphs can substantially facilitate traditionally labor-intensive cross-site comparative analysis in archaeology.

This study has several limitations that warrant further improvement. First, this work only adopts textual materials for model training and benchmark construction, excluding multimodal archaeological evidence such as site photographs, artifact graphics and excavation sketches. Since visual information carries critical contextual and spatial features, incorporating vision-language models in future work can better support comprehensive heritage interpretation. Second, the current knowledge graph is constructed solely based on the Nanlou Site, which limits its generalizability to other archaeological contexts. Future efforts will extend knowledge graph construction to broader site types and develop automated pipelines for cross-site knowledge fusion. Third,

our evaluation partially relies on LLM-based scoring. Although efficient, this method cannot fully replace professional human judgment. Large-scale expert assessment will further enhance the credibility and disciplinary validity of the benchmark. Overall, despite these constraints, this study advances knowledge-grounded AI techniques for archaeological research and provides a feasible bridge between computational tools and traditional heritage scholarship.

Methods

Data collection standards

The collection standards for the archaeological reports are as follows: (1) they must be compiled by official institutions such as national and provincial archaeological research institutes, or the Institute of archaeology at the Chinese academy of social sciences, and should be led or edited by senior experts; (2) priority should be given to sites with significant influence both in China and abroad, covering Chinese history from the Shang and Zhou dynasties through the Ming and Qing periods; (3) coverage should span as many provincial-level administrative regions and major geographical divisions as possible, while encompassing a diverse array of archaeological sub-domains; (4) priority is given to major sites such as World Heritage sites and those recognized as "Top ten new archaeological discoveries in China," and the reports must provide complete and traceable data, including stratigraphy, sketches, and dating results.

The collection standards for the archaeological papers are as follows: (1) include recognized core journals in the field and relevant academic publications from university-affiliated museums; (2) use keywords across 7 major directions such as sites, tombs, and bronzes to focus the topic (e.g., city sites, settlements, casting techniques, and carbon-14 dating); (3) use cultural period terms such as "Shang Dynasty," "Han and Tang Dynasties," and "Liangzhu Culture" to narrow the scope; (4) priority is given to research papers that include experimental data, high-resolution maps and sketches, or well-defined typological and chronological descriptions.

QA pair construction for training

These high-quality, standardized QA pairs can be directly used for fine-tuning large language models. The QA-Ext-Agent fully automates this generation process. Supplementary Fig. 3 illustrates this process, which consists of two main steps: (1) raw

material parsing and multimodal extraction; and (2) multi-agent collaborative generation and quality inspection.

The QA-Ext-Agent first connects to the raw file pool and uses the MinerU[36] tool to automatically parse the input PDF documents in batch. This stage is not limited to plain text extraction; it decomposes the documents into three separate data streams: a collection of images, a JSON-formatted text structure, and a layout file. This detailed preprocessing approach is designed to preserve the relationships between charts and text, thereby providing rich contextual information for downstream QA pair generation.

Subsequently, the system deploys five dedicated agents to complete tasks in a relay manner: Agent 1 (Qwen2.5-72B-Instruct) proposes initial questions; Agent 2 (Qwen2.5-72B-Instruct) refines these questions professionally; Agent 3 (Qwen2.5-72B-Instruct) generates initial answers; Agent 4 (Qwen2.5-72B-Instruct) polishes and improves the answers; and Agent 5 (DeepSeek-R1-Distill-Qwen-32B) performs strict QA matching verification. This setup mimics the rigorous "questioning–writing–proofreading" workflow of human scholars, aiming to mitigate the hallucination issues common in large models and ensure that the generated content can withstand scrutiny in the archaeological context. Meanwhile, three domain experts randomly sample and verify the outputs to ensure the accuracy of the QA pairs. To ensure that the questions extracted by the agents include detailed spatiotemporal information, we designed tailored prompts for them. The detailed prompt design is provided in Supplementary Fig. 4.

After the above verification process, the system saves the qualified QA pairs in json format. This pipeline is capable of large-scale generation, resulting in over 9 million domain-specific QA pairs that have been further processed through back-translation for data augmentation.

Assessment question construction for evaluation

The evaluation question creation process based on the AQ-Gen-Agent is shown in Supplementary Fig. 5 and consists of three main steps: (1) parsing and multimodal decomposition; (2) agent collaboration for question generation and quality inspection; and (3) hierarchical summarization and structured output. In the first stage, the MinerU tool is used to batch-parse PDF documents from the original archaeology corpus, providing a

foundation for subsequent knowledge extraction and question generation.

The second stage involves a pipeline of four dedicated agents: Agent 1 (Qwen3-32B) identifies archaeology knowledge points and predicts their difficulty; Agent 2 (Qwen3-32B) generates questions of various types and cognitive levels based on these knowledge points; Agent 3 (Qwen3-32B) creates reference answers and structured scoring guidelines for each question; and Agent 4 (Qwen3-32B) performs fact-checking, Terminology standardization, and content deduplication. Together, they form a closed-loop quality inspection mechanism that effectively mitigates the risk of large model hallucinations.

Finally, the questions are categorized by difficulty level (basic/intermediate/advanced) and question type, then archived in json format. Each record includes the question text, reference answer, scoring guidelines, difficulty label, archaeology subfield identifier, and source corpus metadata, forming a structured dataset that can be directly used for model evaluation tasks. To ensure that the extracted questions contain rich and accurate spatiotemporal information (such as specific dates, regions, and archaeological cultural layers), each agent is equipped with tailored prompts. The detailed prompt designs are provided in Supplementary Fig. 6.

Implementation detail of model training

The open-source LLMs fine-tuned in this study were implemented using the ModelScope framework, an AI open-source community platform initiated by Alibaba DAMO Academy. The LLM models include Qwen3-8B (https://www.modelscope.cn/models/Qwen/Qwen3-8B), MiniCPM4-8B (https://www.modelscope.cn/models/OpenBMB/MiniCPM4-8B), MiniCPM5-2B (https://www.modelscope.cn/models/OpenBMB/MiniCPM5-2B), Nanbeige4.2-3B (https://www.modelscope.cn/models/nanbeige/Nanbeige4.2-3B), Mistral-7B(https://www.modelscope.cn/models/mlfoundations-dev/mistral_7b_0-3_oh-dcft-v3.1-gpt-4o-mini), ChatGLM3-6B(https://www.modelscope.cn/models/ZhipuAI/chatglm3-6b), and Meta-Llama-3-8B(https://www.modelscope.cn/models/LLM-Research/Meta-Llama-3-8B).

Both LoRA and DoRA fine-tuning was implemented using the PyTorch deep learning framework and the HuggingFace Transformers library[37] . The experiments were conducted

on eight NVIDIA A800 GPUs (80 GB per GPU). We employed the AdamW optimizer with a learning rate of 2×10^{-4}, coupled with a cosine annealing scheduler and a 3% warmup ratio. Training minimized the standard cross-entropy loss for causal language modeling, using a batch size of 32 and a maximum sequence length of 2048 tokens.

For LoRA, we set r=8, alpha=16, and dropout=0.05, applying updates to 7 linear projection layers, resulting in approximately 9.2M trainable parameters (about 0.11% of an 8B model). For DoRA, we set r=16, alpha=32, and dropout=0.1, targeting the same 7 linear projection layers, yielding approximately 9.4M trainable parameters (0.63% of the total). We enabled gradient checkpointing and bf16 mixed precision training, with a gradient clipping threshold of 1.0 and weight decay of 0.01. The data was split into training and validation sets at a ratio of 95%:5%. Model evaluation and checkpoint saving were performed every 100 training steps, and the final model was selected based on the lowest validation loss.

Evaluation of commercial LLMs

The evaluation experiments involving three commercial large language models (DeepSeek-V4-Pro, Qwen3.7-Max, and GLM-5.2) in this study were conducted on the Alibaba Cloud Bailian Large Model Service Platform (https://www.aliyun.com). This platform provides a unified API interface standard supporting high-concurrency request processing, effectively mitigating inference latency fluctuations caused by local hardware differences, thereby ensuring the stability and reproducibility of the evaluation process.

During model interaction, this study employed a structured message-passing mechanism to deliver evaluation questions and instructions to the models. Two evaluation modes were designed, each corresponding to a distinct system prompt configuration: the first mode, "offline without chain-of-thought," configures the model through role assignment in the system prompt and instructs it to directly answer the evaluation questions, outputting only the answer content in a concise and accurate manner with key information included; the second mode, "offline with chain-of-thought," requires the model to reason through three steps: understanding the question, organizing the answer, and providing the answer, and specifies an output format of "[Reasoning Process]...[Final Answer]..." to facilitate separate analysis of the model's reasoning process and final answer in subsequent stages.

Data availability

Our ArchaeoBench dataset will be made publicly available at https://github.com/xxq123sd/kaogushuju.

Code availability

The code for QA-pair generation, model fine-tuning and evaluation is available at https://github.com/xxq123sd/kaogushuju.

Acknowledgements

We thank the archaeology experts who participated in the multi-round review of the QA pairs and assessment questions, and the maintainers of the open-source projects used in this study. This work received no external funding.

Author contributions

K.S. and J.W. conceived and supervised the study. X.X., L.L. and W.Z. designed the dataset construction pipeline and built ArchaeoBench. W.R., F.J., M.C. and B.Q. curated the archaeological corpus and verified QA pairs and assessment questions. X.X., L.L. and D.G. implemented model fine-tuning and evaluation. X.X. drafted the manuscript. Z.J., D.G. and B.Q. provided domain expertise in archaeology. All authors read, revised and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Gattiglia, G. Think big about data: Archaeology and the Big Data challenge. *Archäol. Inf.* **38**, 113–124 (2015).
2. Brandsen, A., Verberne, S., Wansleeben, M. & Lambers, K. Creating a dataset for named entity recognition in the archaeology domain. in *Proc. 12th Conf. Lang. Resour. Eval. (LREC 2020)* 4573–4577 (ELRA, 2020).
3. Baek, S. et al. Single-cell foundation models: bringing artificial intelligence into cell biology. *Exp. Mol. Med.* **57**, 2169–2181 (2025).
4. Liu, F. et al. Application of large language models in medicine. *Nat. Rev. Bioeng.* **3**, 445–464 (2025).
5. Ariai, F., Mackenzie, J. & Demartini, G. Natural language processing for the legal domain: a survey of tasks, datasets, models, and challenges. *ACM Comput. Surv*. 58, 163 (2025).
6. Li, J. et al. An astronomical question answering dataset for evaluating large language models. *Sci. Data* **12**, 447 (2025).
7. Xing, J. et al. Multimodal AI for Yuan Buddhist sculpture chronology and style. *npj Herit. Sci.* **13**, 443 (2025).

8. MiniCPM Team. *MiniCPM4: ultra-efficient LLMs on end devices*. Preprint at https://arxiv.org/abs/2506.07900 (2025).
9. Wang, Y. et al. *Data science and technology towards AGI part I: tiered data management*. Preprint at https://arxiv.org/abs/2602.09003 (2026).
10. Qwen Team. Qwen3 technical report. Preprint at https://arxiv.org/abs/2505.09388 (2025).
11. Llama Team. *The Llama 3 herd of models*. Preprint at https://arxiv.org/abs/2407.21783 (2024).
12. Jiang, A. Q. et al. Mistral 7B. Preprint at https://arxiv.org/abs/2310.06825 (2023).
13. Du, Z. et al. GLM: General language model pretraining with autoregressive blank infilling. in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (ACL 2022)* 320–335 (ACL, 2022).
14. Nanbeige LLM Lab. Nanbeige4.2-3B: unlocking agentic capabilities in a compact model. Preprint at https://arxiv.org/abs/2607.22083 (2026).
15. DeepSeek-AI. DeepSeek-V4: towards highly efficient million-token context intelligence. Preprint at https://arxiv.org/abs/2606.19348 (2026).
16. Cao, Y., Yang, Z. & Wang, Y. Comparative study of large language models on Chinese film script continuation: an empirical analysis based on GPT-5.2 and Qwen-Max. Preprint at https://arxiv.org/abs/2601.14826 (2026).
17. GLM-5 Team. GLM-5: from vibe coding to agentic engineering. Preprint at https://arxiv.org/abs/2602.15763 (2026).
18. Nanjing Museum, Centre for Cultural Heritage and Archaeology, Nanjing University & Jiangyin Museum. *Nanlou: Excavation Report of 2006*. (China Social Sciences Press, 2018).
19. Arslan, M., Ghanem, H., Munawar, S. & Cruz, C. A survey on RAG with LLMs. *Procedia Comput. Sci.* **246**, 3781–3790 (2024).
20. Du, Y. A diachronic study of the Jingnansi settlement during the Xia and Shang periods. in *古代文明（第 15 卷） [Ancient Civilizations, Vol. 15]* (eds 北京大学中国考古学研究中心 & 北京大学震旦古代文明研究中心) 17–29 (上海古籍出版社，上海, 2021).
21. Luoyang Municipal Second Cultural Relics Work Team. The Tang tomb of Lady Qi, née Wu, at Yaling, Yichuan. *Wenwu (*11), 24–44 (1995).
22. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. *In Proc. 10th Int. Conf. Learn. Represent.* (ICLR 2022).
23. Liu, S.-Y. et al. DoRA: weight-decomposed low-rank adaptation. *In Proc. 41st Int. Conf. Mach. Learn.* 32100–32121 (ICML 2024).
24. Wieting, J., Berg-Kirkpatrick, T., Gimpel, K. & Neubig, G. Beyond BLEU: training neural machine translation with semantic similarity. in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.* 4344–4355 (ACL 2019).
25. Karthik, S. A., Ambika, B. J., Naveen, S. & Moorthy, M. U. A framework for automatic summarization of text using text-rank algorithm with ROUGE and BLEU core. in *Data Science and Exploration in Artificial Intelligence: Second International Conference, CODE-AI 2025, Proceedings, Part II* (eds Shreyas, J. et al.) 231–240 (Springer, 2026).
26. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: evaluating text generation with BERT. in *Proc. Int. Conf. Learn. Represent.* (ICLR 2020).
27. Phatthiyaphaibun, W. *et al.* CHIE: generative MRC Evaluation for in-context QA with correctness, helpfulness, irrelevancy, and extraneousness aspects. in *Proc. 2nd GenBench Workshop Gen. (Benchmarking) NLP* 154–164 (ACL 2024).

28. Iminjili, V. et al. A dataset of scientific dates from archaeological sites in eastern Africa spanning 5000 BCE to 1800 CE. *Sci. Data* **12**, 801 (2025).
29. Brugman, H. *et al.* Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC 2016)* 1277–1281 (ELRA, 2016).
30. Bird, D. *et al.* p3k14c, a synthetic global database of archaeological radiocarbon dates. *Sci. Data* **9**, 27 (2022).
31. Li, W., Feng, H., Hu, C., Xu, M. & Cheng, L. Accurate discharge summary generation using fine tuned large language models with self evaluation. *Sci. Rep.* **16**, 5607 (2026).
32. Yeleussinov, A. et al. Fine-tuning Qwen3 models for the legal domain of Kazakhstan: a comparative study of LoRA-adapted models for bilingual legal question answering. *Appl. Sci.* **16**, 6777 (2026).
33. Wang, D., Patel, J., Zha, D., Yang, S. Y. & Liu, X.-Y. *FinLoRA: benchmarking LoRA methods for fine-tuning LLMs on financial datasets*. Preprint at https://arxiv.org/abs/2505.19819 (2025).
34. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst. 35,* 24824–24837 (2022).
35. Hogan, A. *et al.* Knowledge graphs. *ACM Comput. Surv.* **54**, 71 (2021).
36. Wang, B. *et al. MinerU: an open-source solution for precise document content extraction*. Preprint at https://arxiv.org/abs/2409.18839 (2024).
37. Wolf, T. *et al.* Transformers: state-of-the-art natural language processing. in *Proc. 2020 Conf. on Empir. Methods Nat. Lang. Process.: Syst. Demonstr.* 38–45 (ACL 2020).

Supplementary figures

a

INSTRUCTION

ZH 荆南寺遗址西周时期的灰坑H1中发现的陶器主要有哪些?

EN What are the main types of pottery discovered in the ash pit H1 from the Western Zhou period at the Jinnan Temple Site?

OUTPUT

ZH 荆南寺遗址西周时期的灰坑H1中发现的陶器主要器形有鬲、盆、瓮、釜、鼎等，这些陶器反映了西周时期的生活用具。从这些陶器中可以看出，当时居民的日常生活较为丰富，使用了多种不同功能的陶器来满足日常生活需求。比如，鬲用于炊煮，盆用于盛装，瓮用于储物，釜用于烹饪，而鼎则可能是祭祀或重要场合使用的器物。通过对H1出土陶器的研究，不仅能够了解西周时期荆南寺居民的生活方式，还能探讨当时的饮食文化和社会结构。此外，这些陶器的制作工艺和装饰风格，也反映了当时的技术水平和审美倾向，为研究古代社会的发展提供了宝贵资料。

EN The main shapes of the pottery discovered in the ash pit H1 of the Western Zhou period at the Jinnan Temple site include , basins, jars, cauldrons, tripods, etc. These pottery items reflect the daily utensils of the Western Zhou period. From these pottery pieces, it can be seen that the daily life of the residents at that time was relatively rich, and they used various types of pottery with different functions to meet their daily needs. For example, is used for cooking, the basin for holding, the jar for storing things, the cauldron for cooking, and the tripod might be an object used for sacrificial ceremonies or important occasions. Through the study of the unearthed pottery from H1, not only can we understand the lifestyle of the residents of Jingnan Temple during the Western Zhou Dynasty, but also explore the dietary culture and social structure of that time. In addition, the production techniques and decorative styles of these pottery pieces also reflect the technological level and aesthetic tendencies of that time, providing valuable material for the study of cultural development in ancient societies.

b

INSTRUCTION

ZH 齐国太夫人墓出土的玉器中哪些最为精美?

EN Which jade articles unearthed from the tomb of the Grand Duchess of Qi are the most exquisite?

OUTPUT

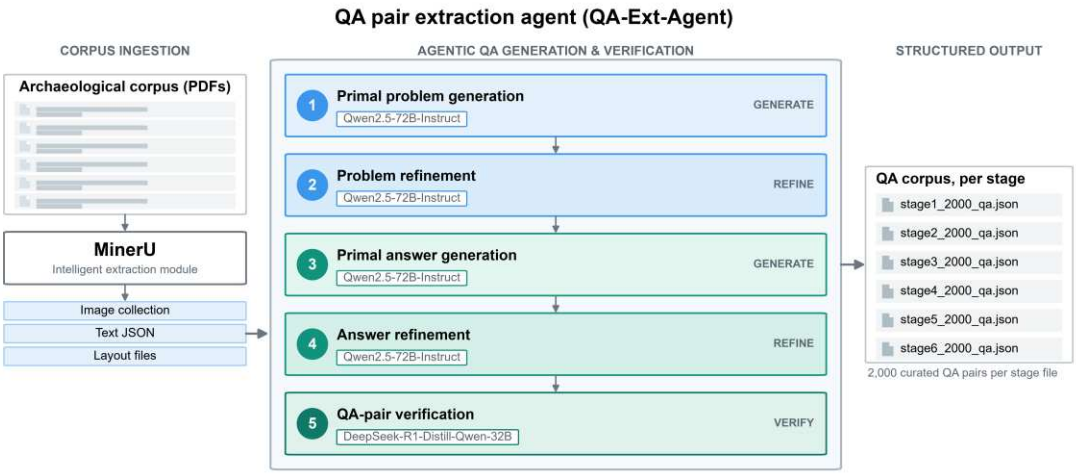
ZH 齐国太夫人墓出土的玉器中最为精美的包括玉钵、玉璜和玉、水晶梳脊。玉钵白玉质，直口，弧腹，圈底，圈足，内壁饰弦纹二周，间以方格纹一周，径8.3厘米、高4厘米，造型典雅，纹饰精美，体现了唐代玉器工艺的高超水平。玉璜共4件，白玉质，形制相同，两端各有一穿孔，径2.1至2.4厘米，造型简约而优雅。水晶梳脊共3件，新月形，底部落有圆形凹槽，用于镶嵌木质梳齿。其中一件白玉质梳脊，两面浅浮雕双凤纹，长7.2厘米、宽7.3厘米，另一件水晶质梳脊，浅浮雕一展翼的蜂蝶，菱形方格纹地，格内填以小团花，边饰以卷云纹，长7.8厘米、宽2.3厘米，工艺考究，形态生动。

EN Among the jade wares unearthed from the tomb of the Matriarch of the State of Qi, the most exquisite ones include jade bowls, jade pendants and jade and crystal combs. The jade bowl is made of white jade, with a straight mouth, an arc-shaped belly, a round bottom and a ring-shaped base. The inner wall is decorated with two strings and one square pattern in between. It measures 8.3 cm in diameter and 4 cm in height. The shape is elegant and the patterns are exquisite, reflecting the superb level of jade craftsmanship in the Tang Dynasty. There are a total of 4 jade pendants, made of white jade and of the same shape. Each end has a perforation with a diameter ranging from 2.1 to 2.4 centimeters. The design is simple yet elegant. There are three pieces of crystal comb spine, crescent-shaped, with circular grooves at the base for inlaying wooden comb teeth. One of the combs is made of white jade, with shallow reliefs of double phoenix patterns on both sides. It measures 7.2 cm in length and 7.3 cm in width. The other is made of crystal, with shallow reliefs of bees and butterflies spreading their wings. It has a diamond-shaped square pattern on the ground, filled with small round flowers inside the grids, and decorated with rolling cloud patterns on the edges. It measures 7.8 cm in length and 2.3 cm in width. The craftsmanship is exquisite and the form is vivid.

Supplementary Fig. 1 Example archaeological QA pairs. a Jingnan Temple Site; **b** Tomb of the Grand Madam of the State of Qi. To minimize semantic ambiguity, each question incorporates comprehensive contextual details, including location, time period, and the specific subject.



Supplementary Fig. 2 Examples of different question types in the ArchaeoBench evaluation dataset. a Choice question; **b** T/F question; **c** Filling in blanks; **d** Terminology; **e** Short Answers. Each question includes a standard answer and specifies the corresponding subfield.



Supplementary Fig. 3 Multi-agent-based generation of high-quality QA pairs. The process is mainly divided into two stages. The first is the intelligent document parsing stage, primarily handled by the MinerU module, which converts unstructured data into structured information. The second is the multi-agent collaborative QA generation stage, which simulates a rigorous academic workflow of question formulation, response drafting, and verification. In this stage, the system deploys five dedicated LLMs to handle tasks sequentially, aiming to mitigate the hallucinations commonly associated with large language models and ensure accuracy in the archaeological context.

a Question generation

```
{
  "role": "A professional text analysis expert, skilled at extracting key information from complex texts and generating structured data that can be used for model fine-tuning (only generating questions).",
  "core_task": "Based on the text provided by the user, generate no less than {number} high-quality questions, which should be described in Chinese.",
  "constraints": [
    "It must be generated directly based on the text content.",
    "The question should have a clear answer orientation.",
    "Different aspects of the text need to be covered.",
    "Ensure the diversity of generated questions.",
    "It is prohibited to generate hypothetical, repetitive or similar questions.",
    "The use of pronouns such as “this”, “that”, “this article”, and “research” is prohibited. The specific object or content they refer to must be clearly defined.",
  ],
  "workflow": [
    "[ Text Analysis ] Process the content in segments, identify key entities and core concepts.",
    "[ Question Generation ] Select the best question point based on information density.",
    "[ Quality Inspection ] Ensure that: The answer can be based on the original text; The tag is strongly relevant to the content of the question. No format errors.",
  ],
  "output_format": {
    "type": "JSON array",
    "example": ["question1", "question2"]
  },
  "requirements": [
    "For names such as personal names and place names, the “English Name (Chinese translation)” format must be used to ensure accuracy and readability.",
  ],
  "note": "It must be output in the prescribed JSON format and no other irrelevant content should be output. Generate no less than {number} high-quality questions; The question should not contain demonstrative pronouns. At least it should include the basic subject, predicate and object."
}
```

b Question semantic completion

```
{
  "role": "Linguists are responsible for supplementing the completeness of questions, ensuring they are unambiguous and have independent and unique identifiers.",
  "core_task": "Based on the source article and the initial question list, generate a more complete question for each question.",
  "framework": "The analytical framework of who (subject), when (time), where (place), why (cause), what (object), and how (method) can be adopted.",
  "completeness_requirements": [
    "It must include the research subject, research object, research time, research location, research purpose, type of technology and related details.",
    "Disambiguation the pronouns (such as “this”, “that”, “it”, “he”, etc.) in the initial question to clarify the specific object or content they refer to.",
  ],
  "constraints": [
    "For each question in the question list, only generate a more complete question. It is prohibited to generate statement statements with periods.",
    "The use of referential words such as “Ben” and “this” is prohibited. When generating, complete entity nouns or contents must be used.",
    "It is prohibited to copy the specific domain keywords in the example (such as MAIER, Turkish Hierapolis of Phrygia, etc.).",
  ],
  "output_format": {
    "type": "JSON array",
    "example": [
      "When the Italian archaeological team conducted on-site analysis of the marble statues unearthed from a certain archaeological site in Turkey in 2014-2015, what non-invasive imaging techniques did the researchers use to detect the traces of ancient paintings?",
    ]
  },
  "note": "For names such as personal names and place names, the “English Name (Chinese translation)” format must be used to ensure accuracy and readability. It is prohibited to generate declarative sentences or expressions with periods."
}
```

c Answer extraction

```
{
  "role": "An expert in fine-tuning dataset generation, skilled at generating accurate question answers from given content.",
  "core_task": "Based on the source text and the list of questions, generate accurate answers related to the questions, totaling {expand question number}.",
  "skills": [
    "The answer must be based on the given content.",
    "The answer must be accurate and must not be fabricated.",
    "The answer must be highly relevant to the question.",
    "The answer must be logical.",
    "If the question is an empty string, then the corresponding answer is also an empty string.",
  ],
  "workflow": [
    "Analyze the given file content step by step.",
    "Extract key information from the content.",
    "Generate accurate answers related to the question.",
    "Ensure the accuracy and relevance of the answers.",
  ],
  "output_format": {
    "type": "JSON array",
    "example": ["question1", "question2"]
  },
  "note": "The answer must be sufficient, detailed, and contain all necessary information, making it suitable for fine-tuning large model training."
}
```

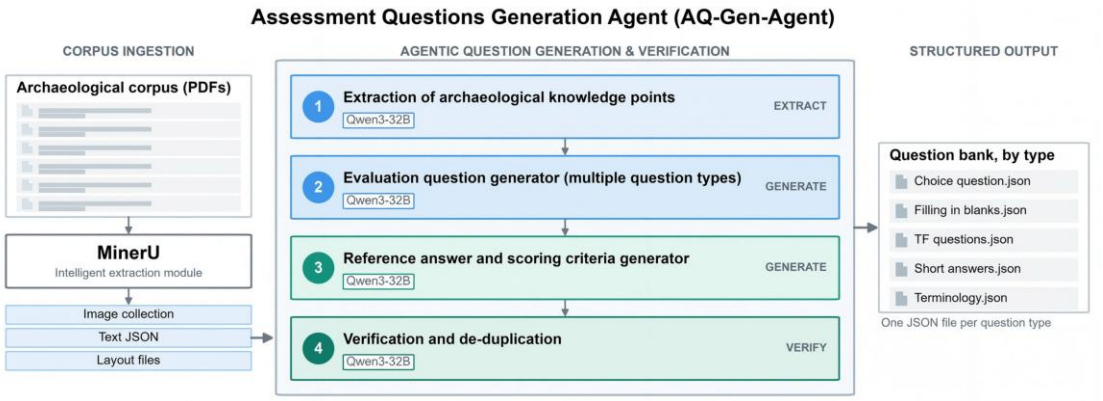
d Answer expansion

```
{
  "role": "An expert in fine-tuning dataset answer expansion, skilled at expanding answers to over 300 characters based on given content and original responses, and ensuring consistency with the original answers.",
  "core_task": "Based on the source text and the original answer list, generate an expanded version for each answer, with a total of {answer number}.",
  "skills": [
    "The answer must be based on the given content.",
    "The answer must be accurate and must not be fabricated.",
    "The answer must be highly relevant to the question.",
    "The answer must be logical.",
    "If the question is an empty string, then the corresponding answer is also an empty string.",
    "The character count of the generated answer must be no less than 300 characters. If it is insufficient, relevant background information can be supplemented, but it is prohibited to directly add questions to the answer.",
  ],
  "workflow": [
    "Analyze the given file content step by step.",
    "Extract key information from the content.",
    "View the corresponding original answer.",
    "Expand each original answer to more than 300 characters.",
  ],
  "output_format": {
    "type": "JSON array",
    "example": ["The expanded answer1", "The expanded answer2"]
  },
  "note": "The answer must be sufficient, detailed, and contain all necessary information, making it suitable for fine-tuning large model training."
}
```

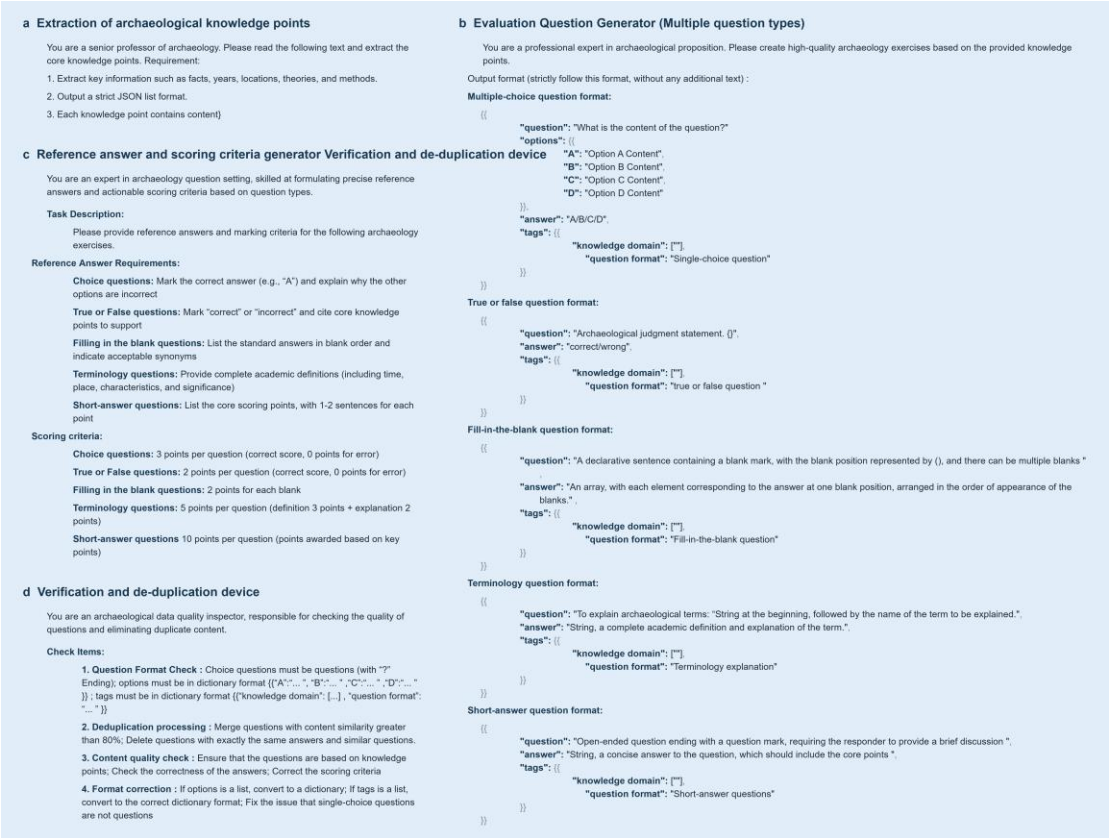
e Answer verification

```
{
  "role": "Senior language experts are responsible for determining whether each answer can correctly address the corresponding question.",
  "core_task": "Based on the source text, question list and answer list, generate a one-dimensional array containing {answer number} Boolean values to represent the validity of each question-answer pair.",
  "workflow": [
    "Summarize and output each question in the question list.",
    "Summarize and output the answer list.",
    "Judge one by one: For each question - answer pair, explain in combination with the source text whether the answer can address the question.",
  ],
  "judgement_criteria": [
    "If the corresponding question or answer is empty, it is false.",
    "It is true only when the answer is accurate, complete and directly related to the question.",
  ],
  "output_format": {
    "type": "JSON array",
    "example": "[true, false]"
  }
}
```

Supplementary Fig. 4 QA pair prompt design. a Question generation; **b** Question semantic completion; **c** Answer extraction; **d** Answer expansion; **e** Answer verification.



Supplementary Fig. 5 Multi-agent-based generation of high-quality evaluation questions. The process can be divided into three main stages: (1) Data parsing and input, in which the MinerU tool converts unstructured data into a machine-readable format; (2) Sequential agent collaboration and generation, in which four specialized LLMs process the data in sequence and form a closed-loop quality inspection mechanism; (3) Structured output and archiving, which produces five types of assessment questions.



Supplementary Fig. 6 Evaluation question prompt design. a Extraction of archaeological knowledge points; **b** Evaluation question generator for multiple question types; **c** Reference answer and scoring criteria generator verification and de-duplication device; **d** Verification and de-duplication device.