ACSE: An Ancient Character Semantic-Aware Embedding for Large Language Models

Zhihan Zhou[1,2], Daqian Shi[2,3], Lida Shi[2,4], Rui Song[1,2], Peiqiang Qiu[2,5], Xiaolei Diao[2,6]*, Hao Xu[1,2]*

[1]College of Computer Science and Technology, Jilin University,
[2]Key Laboratory of Ancient Chinese Script, Culture Relics and Artificial Intelligence, Jilin University,
[3]Digital Environment Research Institute, Queen Mary University of London,
[4]School of Artificial Intelligence, Jilin University, [5]School of Archaeology, Jilin University,
[6]Department of Computer Science, University College London
zhzhou25@mails.jlu.edu.cn, xiaolei.diao@ucl.ac.uk, xuhao@jlu.edu.cn

Abstract

Research on ancient Chinese language is of great significance for tracing Chinese history and civilization. In the field of large language models, studies on the pre-Qin excavated documents such as Oracle Bone Inscriptions, Bronze Inscriptions, and Bamboo Book of Chu remain insufficient. This is because these ancient characters have a low level of digitization, training corpora are extremely scarce, and they typically contain complex and rich semantic information. Therefore, we propose an ancient character semantic-aware embedding for large language models. This embedding integrates both the glyph and lexicality of ancient characters and maps them to the modern Chinese semantic space. We also design a two-stage method for lightweight and parameter-efficient training of the embedding. Finally, we conduct extensive experiments on excavated documents from the pre-Qin period, and the results demonstrate the effectiveness of our approach.

1 Introduction

Research on ancient Chinese language is crucial for reconstructing Chinese history, preserving cultural heritage, and advancing historical linguistics. The recent emergence of large language models (LLMs) has reshaped many areas of natural language processing, yet their ability to comprehend, interpret, and elucidate ancient characters remains limited. Previous work has explored various methods and trained different models for ancient Chinese (Zhang et al., 2024a; Cao et al., 2023; Zhao et al., 2024; Cao et al., 2024), but these approaches largely focus on transmitted ancient Chinese documents. For pre-Qin excavated documents such as Oracle Bone Inscriptions, Bronze Inscriptions, and Bamboo Book of Chu, existing studies are still insufficient (Shi et al., 2022b; Diao et al., 2023; Guiyuan, 2023).

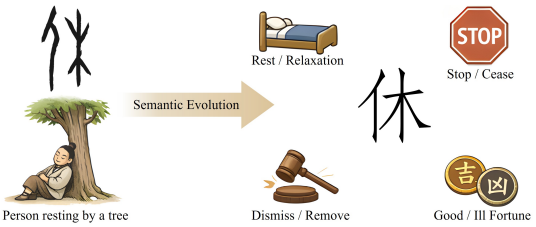

Figure 1: An example of an ancient character containing rich semantic information. The oracle bone script "rest" carries meanings associated with its component parts, while also developing new meanings across different media and historical periods as it evolved over time.

The scientific problems presented by the pre-Qin excavated documents make LLM training especially challenging (Diao et al., 2025; Yue et al., 2025; Zhou et al., 2026). These ancient characters are sparsely digitized, and many are not encoded in contemporary computing standards. Although (Diao et al., 2026) proposes a practical pipeline for digitizing ancient characters, the problem of how to integrate encoded ancient characters into the representational vocabulary of LLMs remains unresolved. Existing methods can add new token indices and fine-tune the model (Kajiura et al., 2023; Kim et al., 2024; Yamaguchi et al., 2025; Takase et al., 2025; Sha et al., 2025; Han et al., 2025), but assigning indices for ancient characters entails initializing new embeddings, which typically rely on hand-crafted heuristics and therefore require substantial data in the domain to learn high-quality representations. The pre-Qin excavated corpora are far too limited to meet these requirements. In addition, ancient characters carry rich, complex semantic information that standard fine-tuning often fails to capture: many ancient graphemes are pictographic, their component parts themselves encoding meaning; over time, the senses of these graphemes shift

*Corresponding Authors

Findings of the Association for Computational Linguistics: ACL 2026, pages 9000–9012

across media and historical periods. As shown in the Figure 1, on the left is the character "rest" in oracle bone script, whose glyph resembles a person leaning against a tree, depicting the action scene of "resting or taking a break". On the right is the modern Chinese character "rest", which has evolved over historical periods to meaning "rest", "cease", "dismiss", and "good/ill fortune" in divination. LLMs need to comprehend both glyph information and the lexicality across different historical periods to achieve the understanding and representation of ancient characters.

Extensive research indicates that the meanings of ancient Chinese characters can be traced (Boltz, 1986; Jane, 2014). Modern Chinese characters are the product of long historical evolution, and many semantic cues preserved in ancient characters remain, directly or indirectly, in modern radicals, lexical patterns, and semantic distributions (Chi et al., 2022; Shi et al., 2022a). LLMs already capture deep semantic and syntactic regularities of modern Chinese. If one can build a robust mapping from ancient graphemes into the semantic space of modern Chinese encoded, then targeted adaptation using limited ancient corpora can substantially improve the model's capacity to interpret ancient Chinese documents. The central idea is to exploit shared representations rather than relying on massive new pretraining. By taking advantage of continuity between ancient and modern writing systems, a modern LLM can be nudged to understand ancient characters with far fewer data.

In this work, we propose a method based on Qwen2.5 that enhances the glyph and semantic information of ancient Chinese characters. Under low-resource constraints, we adopt two practical design choices. First, The byte-level byte-pair encoding (BBPE) tokenizer encodes text at byte granularity and can represent both modern and ancient characters without expanding the model's token index space. In other words, prior to any adaptation, ancient and modern characters share a common byte space, which substantially reduces the need to create and train new token embeddings from scratch. By exploiting this shared byte-level tokenization, we obtain initial implicit associations between ancient and modern glyphs that require far less data to fine-tuning than learning new tokens. Second, for ancient character embeddings, we propose an ancient character semantic-aware embedding that models two uniquely important features, namely glyph and lexicality. Our embedding merges both glyph and lexical embeddings and then projects the fused features into the modern Chinese semantic space. Glyph features help the model recover component-anchored semantic cues, while lexical features supply the sense information that emerges through historical change.

For training we adopt a multi-stage, parameter-efficient adaptation strategy applied to the Qwen2.5 backbone. Rather than densely fine-tuning the entire model, we perform light, targeted fine-tuning using contrastive learning and instruction fine-tuning to train the large language model. The goal is to inject glyph and lexical priors that guide semantic hypotheses under data scarcity, and to map inferred ancient senses into the existing modern Chinese representational space so that the language model requires minimal parameter adjustment. This design acknowledges both the reality of scarce archaeological resources and maximally leverages the knowledge already encoded in modern LLMs. Our contribution is summarized below.

- We propose an ancient character semantic-aware embedding. This method integrates both glyph and lexical information of ancient Chinese characters and projects them into the semantic space of modern Chinese.
- We propose a two-stage training method. We first inject prior glyph and lexical knowledge, then guide the model to formulate word-meaning hypotheses and perform inference, thereby maximizing the utilization of knowledge inherently encoded within modern large language models.
- We design multiple evaluation tasks based on pre-Qin excavated documents and conduct extensive experiments. The results demonstrate the effectiveness of our approach.

2 Related Work

The development of historical linguistics and the preservation of cultural heritage have consistently been significant research topics in the field of natural language processing. (Assael et al., 2022) developed Ithaca for ancient Greek inscriptions. (Son et al., 2022) proposed HUE, capable of translating Old Korean script into modern Korean and English. In the field of ancient Chinese, early research primarily focused on training models for specific tasks (Shi et al., 2025a,b). (Han et al., 2018) performed named entity recognition on ancient Chi-

nese corpora, while (Chang et al., 2021; Jiang et al., 2023) conducted translation tasks within this domain. However, most of these approaches required reliance on vast amounts of manually annotated data to achieve satisfactory performance. With the rise of pre-trained language models (PLMs), numerous researchers have extended their work to models such as BERT and GPT (Yu and Wang, 2020; Tian et al., 2021; Wang et al., 2022; Chang et al., 2023; Yao et al., 2025). In recent years, LLMs have been introduced into the field of ancient Chinese because of their exceptional inductive and reasoning capabilities, giving rise to numerous evaluation benchmarks (Li et al., 2021; Yao et al., 2022; Pan et al., 2022; Zhang and Li, 2023; Wei et al., 2024; Diao, 2022; Giunchiglia et al., 2023) and LLMs (Zhang et al., 2024a; Cao et al., 2023; Zhang et al., 2024b). Xunzi (Zhao et al., 2024) was trained on ancient Chinese corpora and demonstrated outstanding performance in lexical analysis and contextual comprehension. Tonggu (Cao et al., 2024) achieved excellent performance across multiple ancient Chinese tasks. However, existing research has largely focused on transmitted documents, with scant attention paid to pre-Qin excavated documents. AncientBench (Zhou et al., 2026) bridges this gap by proposing a benchmark tailored to the context of pre-Qin excavated texts. We propose ACSE, which is based on the BBPE tokenizer and models two features, through a two-stage training approach, thereby significantly enhancing the LLM's comprehension of pre-Qin excavated documents.

3 Background and Preliminary Study

We select Qwen2.5 as the backbone of our approach primarily because its BBPE tokenizer offers a practical path for low-resource adaptation. Unlike character- or word-level tokenizers, BBPE operates directly on UTF-8 encoding i.e., the 1–4 byte sequences that encode any Unicode code point. Consequently, any modern or ancient glyph expressible in Unicode can be represented without requiring additional token indices. During training, the BBPE tokenizer learns high-frequency byte sequences that may correspond to whole characters, character fragments, or multi-character spans, and maps those sequences to indices in a fixed vocabulary. These design choices give BBPE tokenizer a shared index space that implicitly embeds many modern and ancient features while avoiding the

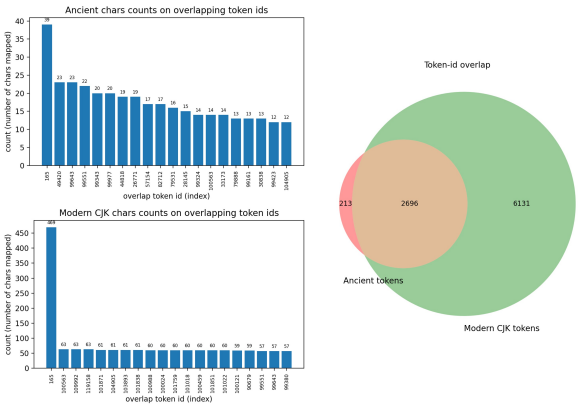

Figure 2: Display of index sharing between modern and archaic characters in the BBPE tokenizer. The two left figures show the top 20 overlapping indices occurring most frequently in the ancient and modern character sets respectively. The horizontal axis represents overlapping indices, while the vertical axis indicates the frequency of each index within the character set. The right figure provides a visual representation of overlapping indices between ancient and modern characters.

costs of expanding and re-training a character-level vocabulary. However, directly sharing indices is only part of the solution, as BBPE disperses semantic signals across multiple bytes or sub-token units and is trained on modern Chinese, it must be supplemented by fusion training that combines both glyph and lexicality.

To quantify the degree of index sharing between modern and ancient characters in the BBPE tokenizer, we conduct a preliminary index overlap study. We employ the Qwen2.5 tokenizer to process both modern (CJK unified ideographs U+4E00-U+9FFF) and ancient Chinese character corpora, recording the tokenization index for each character and measuring the intersection between the two character sets, the results are as shown in the Figure 2. In the figures on the left, the most frequently occurring character indices in both the ancient and modern Chinese character sets are 165 index. Indices such as 99551, 99643, and 100563 are among the top 20 most frequent indices in both ancient and modern Chinese character sets. In the figure on the right, we observe that the ancient character set and modern character set comprise 2,909 and 8,827 characters respectively, with 2,696 characters exhibiting index overlap. These preliminary experiments demonstrate that index sharing between modern and ancient characters is highly prevalent within the BBPE tokenizer.

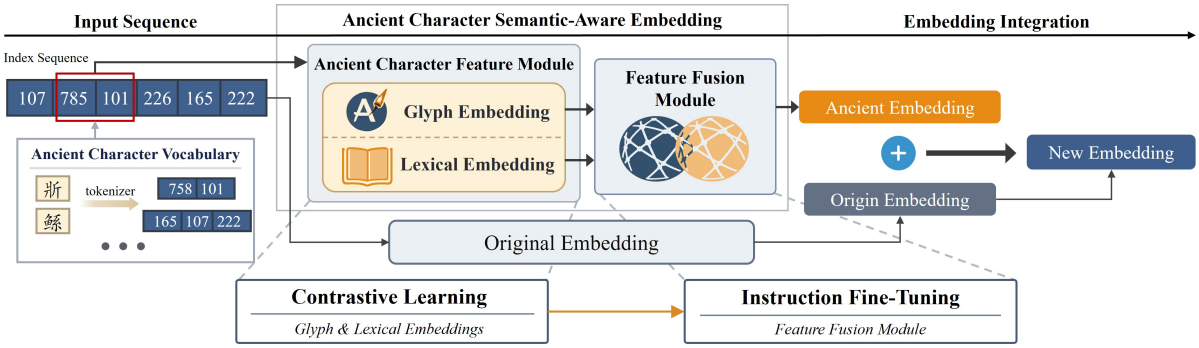

Figure 3: An overview of ancient character semantic-aware embedding. The sub-sequences within the model's input index sequence that appear in the ancient character vocabulary are fed into the ancient character feature module to obtain glyph embedding and lexical embedding, then into the feature fusion module to obtain the ancient embedding. The ancient embedding is added and concatenated with the origin embedding to produce the new embedding. We conduct contrastive learning on glyph/lexical embedding and instruction fine-tuning on the feature fusion module.

4 The Proposed Method

We propose an ancient character semantic-aware embedding method. Specifically, we model two features of unique significance to ancient characters to participate in the inference of the model (§4.1), and conduct training for the ancient character semantic-aware embedding using a two-stage training approach (§4.2).

Many ancient Chinese scripts, including Oracle Bone Inscriptions, Bronze Inscriptions, and Bamboo Book of Chu, remain unencoded in computer systems because of the remoteness of their historical era and the scarcity of reference materials, i.e., numerous ancient characters are excluded from the Unicode encoding system. Consequently, we employ InteChar (Diao et al., 2026), a character list integrating previously unencoded oracle bone characters alongside traditional and modern Chinese characters as our ancient character vocabulary V^{char}. The construction process of InteChar is detailed in Appendix A.

4.1 Ancient Character Semantic-Aware Embedding

After obtaining the ancient character vocabulary V^{char}, we employ the Qwen2.5 tokenizer to encode all ancient characters. As the Qwen2.5 tokenizer is based on BBPE encoding, all ancient characters are encoded into an ancient index subsequence, producing an ancient character index vocabulary V^a.

$$V^a = tokenizer(V^{char}) \tag{1}$$

We then construct a new embedding that incorporates two features of unique significance for ancient characters, integrating both into the model's reasoning process. We name this new embedding ancient character semantic-aware embedding (ACSE). ACSE comprises two modules, i.e., the ancient character feature module and the feature fusion module, which we will introduce in detail below.

Ancient Character Feature Module. In the ancient character feature module, we define two embeddings: glyph embedding $E^{glyph} \in \mathbb{R}^{|V^{char}| \times h}$ and lexical embedding $E^{lexical} \in \mathbb{R}^{|V^{char}| \times h}$, where h is an embedding dimension. Upon receiving a sentence as input $S^{input} : t \rightarrow (t_1, t_2, ..., t_k)$, the model processes it through the tokenizer to obtain an index sequence $S^w : w \rightarrow (w_1, w_2, ..., w_k)$, Where t and w denote each token and index in S^{input} and S^w respectively.

$$S^w = tokenizer(S^{input}) \tag{2}$$

The ancient character feature module then traverses this index sequence. If any sub-sequence is contained within the ancient character index vocabulary i.e., the input sentence contains ancient characters, that sub-sequence is fed into the glyph embedding and lexical embedding. To match the shape of the original embedding, we multiply the glyph embedding and lexical embedding by the $\mathbf{1}_{|x|} = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{|x| \times 1}$ matrix respectively and take the average, where x denotes the index of the sub-sequence. E^f denotes the set of all ancient sub-sequence features. The calculation process is as follows.

$$S^x = \{S^w_{1,2}, ..., S^w_{i,j}\}, 0 \leq i < j \leq |S^w|) \tag{3}$$

$$E(x) = \frac{1}{|x|}\mathbf{1}_{|x|} E_{\mathcal{I}(x)}, x \in S^x \cap V^a \tag{4}$$

$$E^f = \{E(x_1), ...E(x_m)\}, x_m \in S^x \cap V^a \quad (5)$$

where $S^w_{i,j}$ denotes the index sequence of S^w from position i to j, and S^x denotes all sub-sequences of S^w. $E \in \{E^{glyph}, E^{lexical}\}$ denotes the glyph embedding and lexical embedding, $\mathcal{I}(x)$ denotes the index of subsequence x in V^a, $E(x)$ denotes the processing of individual x, and E^f denotes the set of all ancient subsequence features.

Feature Fusion Module. Within the feature fusion module, we combine the origin embedding of the index input to the model with the ancient text embedding obtained from the ancient character feature module to generate a new input embedding. This is then fed into subsequent layers of the model for inference. The calculation process is as follows.

$$E^c_x = concat(E^{glyph}_x, E^{lexical}_x), x \in S^x \cap V^a \quad (6)$$

$$E^{anc}_x = \alpha \cdot (WE^c_x + b) \quad (7)$$

$$E^{new} = \begin{cases} E^o_x + E^{anc}_x, & s \in S^x \cap V^a \\ E^o_x, & s \in S^x \setminus V^a \end{cases} \quad (8)$$

Where E^{glyph}_x and $E^{lexical}_x$ perform the operation on $E(x)$ in formula (4), E^c_x denotes the concatenated embedding. To ensure the embeddings share the same shape, we process E^c_x through a linear layer, with W and b as learnable vectors and α as a learnable parameter. Finally, we fuse E^{anc}_x with the original embedding to calculate E^{new}, where E^o_x denotes the original embedding.

4.2 A Two-Stage Training Method

We conduct lightweight, parameter-efficient adaptive fine-tuning. we adopt a two-stage training approach because of the scarcity of available training corpora for pre-Qin excavated documents, i.e., first injecting prior knowledge based on character glyph and lexicality, then training the model to fuse ancient script feature information and perform inference.

Training of Ancient Character Feature Module. In the first stage, we infuse ancient text feature information. Based on the ancient knowledge graph (Diao et al., 2026) and the ancient Chinese dictionary ShuoWen (Xu, 1981), we build a glyph dataset and a lexical dataset for training glyph embeddings and lexical embeddings respectively. The data format consists of pairs of ancient characters and their corresponding glyphs/lexicality. Our training objective is to align ancient characters with their relevant glyphs/lexicality while keeping them distant from unrelated glyphs/lexicality. We employ a contrastive learning approach, using InfoNCE (Rusak et al., 2024) as the loss function. The calculation process is as follows.

$$E^{char} = E(c), c \in V^a \cap D \quad (9)$$

$$E^{pos} = E^{original}_{pairs}, pairs \in D \quad (10)$$

$$s = \sum_{t=1}^{K} e_t, p = \frac{s}{K}, e_t \in \{E^{char}, E^{pos}\} \quad (11)$$

$$N = \frac{p}{||p||_2 + \epsilon} = \frac{s}{||s||_2 + K\epsilon} \quad (12)$$

$$q = N(e), e \in E^{char}, k = N(e), e \in E^{pos}, \quad (13)$$

$$\mathcal{L}^{emb} = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{exp(\frac{q_i \cdot k_i+}{\tau})}{\sum_{j=1}^{N} exp(\frac{q_i \cdot k_i-}{\tau})}) \quad (14)$$

Where D denotes the glyph and lexical datasets, E denotes glyph embedding and lexical embedding, and ϵ and τ are hyperparameters. We first calculate the ancient embedding for the ancient characters and the original embedding for the glyph/lexicality. To ensure training stability, we normalize the embeddings to obtain q and k, representing the query sample and key sample respectively. Then we calculate the infoNCE loss, where k_i+ denotes positive samples and k_i- denotes negative samples. We define all samples except those corresponding to the query sample as negative samples. More details on training of ancient character feature module can be found in Appendix B.

Training of Feature Fusion Module. In the second stage, we train the model to integrate ancient feature and perform reasoning. We construct an ancient instruction fine-tuning dataset, loading the glyph embedding and lexical embedding trained in the first stage into the model. We fix the model parameters and glyph/lexical embedding parameters, training only the linear layers and α parameters, performing end-to-end instruction fine-tuning. The loss function is as follows, where $M = \sum_{i,t} m_{i,t}$.

$$\mathcal{L}^{sft} = -\frac{1}{M}\sum_{i,t} m_{i,t} log p_\theta(y_{i,t}|x_i, y_{i,<t}) \quad (15)$$

5 Experiments

In this section, we conduct extensive experiments to validate the effectiveness of our approach. We first evaluate ancient character comprehension, demonstrating that ACSE enhances model understanding and reasoning capabilities within ancient

Model	Radical		Radical Meaning		ExcDoc Word		Phonetic Loan Character	
	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot
Qwen-7B-Chat(Bai et al., 2023)	45.62	43.68	45.99	41.38	69.32	71.51	35.17	37.74
Llama3-8B-Instruct(Grattafiori et al., 2024)	42.55	43.96	40.54	38.52	70.96	74.52	47.85	53.27
Baichuan2-7B-Chat(Yang et al., 2025a)	50.36	40.69	48.29	44.80	70.68	70.68	42.01	39.10
Yi1.5-9B-Chat(Young et al., 2025)	42.49	45.31	42.92	48.15	75.89	78.90	44.73	51.91
Baichuan2-13B-Chat(Yang et al., 2025a)	48.09	44.21	42.08	43.48	75.34	73.70	**56.72**	48.76
GLM4-9b-chat(Zeng et al., 2024)	47.21	50.87	42.71	41.80	**81.37**	81.37	42.16	46.88
Qwen-14B-Chat(Bai et al., 2023)	**53.05**	47.67	**52.97**	46.96	74.79	76.71	54.86	58.77
Qwen2.5-7B-Instruct(Yang et al., 2025b)	45.55	**55.55**	45.99	**50.70**	80.27	**82.42***	53.96	**60.66**
Tonggu-7b-chat(Cao et al., 2024)	36.75	34.97	30.70	34.12	57.26	63.29	28.68	29.65
Xunzi-Qwen-Chat(Zhao et al., 2024)	40.08	**45.65**	45.99	43.61	71.23	70.96	36.12	39.83
Yi1.5-9B-Ancient(Zhou et al., 2026)	**52.42**	45.55	**49.13**	**51.36**	**77.26**	**80.27**	**40.80**	**51.18**
ACSE-Qwen2.5-Instruct	62.02*	70.01*	66.53*	72.61*	80.52*	81.91	65.87*	77.94*

Table 1: Evaluation results for the model dimensions of ancient character comprehension. The top part shows the modern Chinese model, the middle part the ancient Chinese model, and the bottom part the Qwen2.5-7B-Instruct model fine-tuned using the ACSE method. Bold indicates the highest score for each task in the modern Chinese and ancient Chinese models respectively, while "*" denotes the highest score for each task.

script scenarios. Subsequently, we visually illustrate the representation of ACSE, confirming its ability to map ancient characters to appropriate positions within representation spaces. Finally, we conduct ablation experiments on modules of ACSE, verifying the necessity of each module within the architecture.

5.1 Setup

Datasets. In the evaluation of ancient character comprehension, we employ AncientBench (Zhou et al., 2026) as our dataset, selecting tasks closely related to excavated ancient documents including Radical, Radical Meaning, ExcDoc Word, and Phonetic Loan Character. For the ancient character representation experiments, we construct a dataset of ancient characters paired with their glyph/lexicality based on the Radical task source data and ShuoWen (Xu, 1981) to validate model performance.

Baselines. All evaluations are conducted on 11 models and 4 fine-tuning approaches. We select 8 LLMs with strong Chinese capabilities, including Qwen-7B-Chat (Bai et al., 2023), Llama3-8B-Instruct (Grattafiori et al., 2024), Baichuan2-7B-Chat (Yang et al., 2025a), Yi1.5-9B-Chat (Young et al., 2025), Baichuan2-13B-Chat (Yang et al., 2025a), GLM4-9b-chat (Zeng et al., 2024), Qwen-14B-Chat (Bai et al., 2023), and Qwen2.5-7B-Instruct (Yang et al., 2025b), 3 ancient Chinese LLMs, including Tonggu-7b-chat (Cao et al., 2024), Xunzi-Qwen-Chat (Zhao et al., 2024), and Yi1.5-9B-Ancient (Zhou et al., 2026), and 4 fine-tuning methods, including Prefix (Li and Liang, 2021), Adapter (Houlsby et al., 2019), BitFit (Ben Zaken et al., 2022), and LoRA (Hu et al., 2021).

Implementation details. For training, we employ Qwen2.5-7B-Instruct as our foundational model. During training of the ancient character feature module, we set the batch size to 16, learning rate to 3e-4, and hyperparameter τ to 0.1, running for 12 epochs. For the feature fusion module training, we set the batch size to 2, learning rate to 1e-3, and employ the AdamW optimizer, fixing all other model parameters while training only the linear and α for 3 epochs. All experiments are executed on the machine running the Ubuntu OS with ascend-910b npu. For evaluation of ancient character comprehension, we record the model's accuracy under 0-shot and 5-shot conditions. For evaluation of ancient character representation capability, we visualize the positional relationships between ancient characters and their glyphs/lexicalitys by mapping them into a 2D distribution using multidimensional scaling (MDS). More implementation details can be found in Appendix C.

5.2 Ancient Character Comprehension Evaluation

We select 4 tasks from Ancient Bench that are highly relevant to excavated texts: Radical and Radical Meaning are relevant to glyph competence, ExcDoc Word is relevant to lexical meaning competence, and Phonetic Loan Character is relevant to contextual competence. We evaluate ACSE in two dimensions: the model and the fine-tuning method.

Dimension of Models. We conduct evaluations on 11 models with strong Chinese language capabilities, encompassing both modern language

Method	Radical		Radical Meaning		ExcDoc Word		Phonetic Loan Character	
	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot
Prefix(Li and Liang, 2021)	26.20	26.57	25.30	25.70	26.17	26.17	24.83	25.54
Adapter(Houlsby et al., 2019)	26.74	26.84	23.11	23.11	28.12	28.12	25.32	26.16
BitFit(Ben Zaken et al., 2022)	44.53	53.44	45.82	50.70	83.20*	81.25	52.00	58.13
Origin(Yang et al., 2025b)	45.55	55.55	45.99	50.70	80.27	82.42	53.96	60.66
LoRA(Hu et al., 2021)	44.74	55.26	46.55	51.10	78.63	82.42*	39.49	60.91
ACSE	62.02*	70.01*	66.53*	72.61*	80.52	81.91	65.87*	77.94*

Table 2: Evaluation results for the fine-tuning methods dimensions of ancient character comprehension. "*" denotes the highest score for each task.

models and classical Chinese models, with results presented in the Table 1. ACSE-Qwen2.5-Instruct demonstrates outstanding performance across all four tasks, achieving the highest scores in Radical, Radical Meaning, and Phonetic Loan Character tasks. In the Radical and Radical Meaning tasks, ACSE-Qwen2.5-Instruct achieves 5-shot scores of 70.01 and 72.61 respectively, surpassing the current state-of-the-art models by 14.46 and 21.25. This improvement stems from the fact that existing LLMs' tokenization processes largely exclude characters from pre-Qin excavated documents such as Oracle Bone Inscriptions, Bronze Inscriptions, and Bamboo Book of Chu. The ancient vocabulary and ancient character feature modules of ACSE compensate for this information gap, enabling ACSE-Qwen2.5-Instruct to deliver superior performance. It is worth noting that ACSE-Qwen2.5-Instruct also achieves outstanding performance in the Radical Meaning task. This task requires models not only to understand ancient character components, but also to comprehend the meanings of these components. Consequently, the score of ACSE-Qwen2.5-Instruct in the Radical Meaning task further validates ACSE's capability in ancient character feature fusion and its ability to map ancient characters to modern Chinese characters. ACSE-Qwen2.5-Instruct achieves a 5-shot score of 77.94 on the Phonetic Loan Character task, surpassing the current state-of-the-art model by 17.28 points. This demonstrates the Feature Fusion Module's enhancement of both feature integration and downstream task reasoning capabilities. ACSE-Qwen2.5-Instruct achieves a relatively strong score of 81.91 on the 5-shot ExcDoc Word task. We also evaluate its performance in 0-shot scenes, observing that 5-shot scores consistently outperformed 0-shot scores across all tasks, a pattern not necessarily observed in baseline models. This discrepancy likely stems from the differing formats of the instruction fine-tuning task (instruction-based question answering) and the evaluation task (multiple-choice questions), necessitating the inclusion of a small number of examples as reference during assessment. This further illustrates how task format gaps can degrade model performance, indicating that in practical applications, providing minimal prompts or conducting further fine-tuning is required to fully stimulate the capabilities of ACSE-Qwen2.5-Instruct.

Dimension of Fine-Tuning Methods. To demonstrate the efficacy of ACSE, we fine-tune Qwen2.5-7B-Instruct using diverse methods. As no existing fine-tuning approach specifically targets pre-Qin excavated documents, we compare ACSE against the four most prevalent fine-tuning methods, with results presented in the Table 2. ACSE achieves the best performance across nearly all tasks. The most significant improvement is observed in the two glyph competence tasks, consistent with our aforementioned explanation that the ancient character feature module provides glyph information. It is worth noting that both Adapter and Prefix approaches resulted in substantial negative performance shifts, even causing scores to approach 25% across four tasks, exhibiting random guessing behavior. We hypothesize this occurs because new and old characters share identical indices. During fine-tuning, newly acquired knowledge may disrupt the original representations, and without appropriate structural guidance to harness this knowledge effectively, the representation space may become corrupted or even lose its original textual representations. These experimental results further demonstrate that traditional fine-tuning methods are not fully applicable to the context of pre-Qin excavated texts and may potentially damage the model's original representation space.

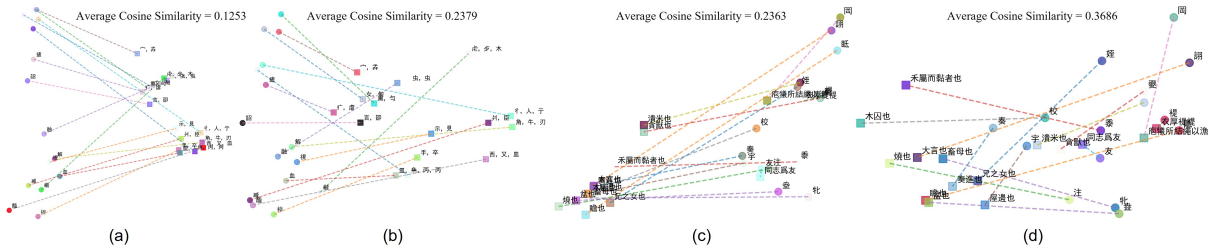

Figure 4: Visualisation results of the representation space. (a) and (b) indicate the representation results of ancient character glyph information in the original embedding and ACSE respectively. (c) and (d) indicate the representation results of ancient character lexical information in the original embedding and ACSE respectively.

w/o embedding	Rad	R.M	Exc	P.L.C
glyph	54.13	50.00	80.86	65.16
lexical	54.18	49.90	80.86	65.25
ACSE	70.01	72.61	81.91	77.94

Table 3: The ablation study of ancient character feature module, where "glyph" and "lexical" denote results obtained by removing glyph embedding and lexical embedding respectively.

5.3 Visualization of Representation Space

To further observe the representation of ancient characters by ACSE, we visualize the relative positions of ancient characters and their glyphs/lexicality within the representation space. We employ cosine similarity as the metric for relative positioning, as this more accurately reflects the semantic relationships between characters or words. We projected the relationships between ancient characters and character glyphs/lexicality pairs into 2D space, with the results depicted in the Figure 4, where points represent ancient characters, glyphs, or lexicalities, and dashed lines denote relationships. Closer distances indicate higher cosine similarity and greater semantic proximity, with closer distances indicating higher cosine similarity and greater semantic proximity. In the figures of the relationship between ancient characters and glyphs, figure (a) shows the original embedding representation, with ancient characters and glyphs distributed on opposite sides. Ancient characters exhibit a more dispersed distribution, while glyphs cluster more densely. This aligns with our intuition, as the original embedding lacks prior information about ancient characters, treating them as unknown symbols, whereas most glyphs are contained within modern Chinese characters. Consequently, ancient characters and glyphs are relatively distant, with ancient characters distributed more sparsely. Figure (b) displays the ACSE representation. Following contrastive learning and instruction fine-tuning, the distance between ancient characters and their corresponding glyphs has been significantly reduced, and the distribution of ancient characters and glyphs has become more uniform. In the figures of the relationship between ancient characters and meanings, Figure (c) shows the original representation, with meanings and ancient characters distributed on opposite sides. Figure (d) presents the ACSE representation, where the trained representation exhibits a more uniform distribution.

5.4 Ablation Study

To further validate the effectiveness of each modules in ACSE, we conduct ablation experiments on the ancient character feature module. We separately record ACSE's experimental results under 5-shot conditions after removing either the glyph embedding or lexical embedding. The experimental results are shown in Table 3. For tabular clarity, we use Rad, R.M, Exc, and P.L.C to represent the four tasks: Radical, Radical Meaning, ExcDoc Word, and Phonetic Loan Character respectively. The experimental results demonstrate that the absence of any module significantly impacts ACSE's performance. In the Radical Meaning task, the absence of lexical embedding caused a 22.71 decline in ACSE performance. Overall, the impact of omitting glyph and lexical embeddings is similar for ACSE, with the loss of either embedding reducing performance to levels comparable to original embedding or LoRA fine-tuning methods. These results demonstrate the efficacy and irreplaceable nature of glyph and lexical embedding.

6 Conclusion

In this study, we propose an ancient character semantic-aware embedding that models two features of unique significance to pre-Qin excavated

documents and performs two-stage training to achieve efficient understanding of such documents by large language models. Extensive experiments across multiple dimensions and formats demonstrate the effectiveness of our approach.

Limitations

In this paper, we propose an ancient character semantic-aware embedding for LLMs, enabling LLMs to comprehend pre-Qin excavated documents. However, current approaches remain confined to textual modalities for capturing semantic information, necessitating the incorporation of additional modalities to enhance LLM comprehension of excavated documents. Furthermore, our experiments are primarily conducted on Qwen2.5, with its efficacy yet to be explored across more BBPE-based models. Future research will focus on addressing these limitations and exploring the application potential of multimodal models.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62476111), the Department of Science and Technology of Jilin Province, China (20230201086GX), the "Paleography and Chinese Civilization Inheritance and Development Program" Collaborative Innovation Platform (No.G3829), the National Social Science Foundation of China (No. 23VRC033), and the interdisciplinary cultivation project for young teachers and students at Jilin University, China (No. 2024-JCXK-04).

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando De Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and Jianxin Ma. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1–9. Association for Computational Linguistics.

William G Boltz. 1986. Early chinese writing. *World Archaeology*, 17(3):420–436.

Jiahuan Cao, Dezhi Peng, Yongxin Shi, Zongyuan Jiang, and Lianwen Jin. 2023. Translating ancient chinese to modern chinese at scale: A large language model-based approach. In *Proceedings of ALT2023: Ancient Language Translation Workshop*, pages 61–69.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. Tonggu: Mastering classical chinese understanding with knowledge-grounded large language models. *Preprint*, arXiv:2407.03937.

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. 2021. Time-aware ancient chinese text translation and inference. *arXiv preprint arXiv:2107.03179*.

Liu Chang, Wang Dongbo, Zhao Zhixiao, Hu Die, Wu Mengcheng, Lin Litao, Shen Si, Li Bin, Liu Jiangfeng, Zhang Hai, and 1 others. 2023. Sikugpt: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *arXiv preprint arXiv:2304.07778*.

Yang Chi, Fausto Giunchiglia, Daqian Shi, Xiaolei Diao, Chuntao Li, and Hao Xu. 2022. Zinet: Linking chinese characters spanning three thousand years. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3061–3070.

Xiaolei Diao. 2022. Building a visual semantics aware object hierarchy. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, IJCAI-ECAI 2022*.

Xiaolei Diao, Daqian Shi, Wei Cao, Ting Wang, Ruihua Qi, Chuntao Li, and Hao Xu. 2025. Oracle bone inscription image restoration via glyph extraction. *npj Heritage Science*, page 321.

Xiaolei Diao, Daqian Shi, Jian Li, Lida Shi, Mingzhe Yue, Ruihua Qi, Chuntao Li, and Hao Xu. 2023. Toward zero-shot character recognition: A gold standard dataset with radical-level annotations. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6869–6877.

Xiaolei Diao, Zhihan Zhou, Lida Shi, Ting Wang, Ruihua Qi, Daqian Shi, and Hao Xu. 2026. Intechar: A unified oracle bone character list for ancient chinese language modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 211–219.

Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao. 2023. Aligning visual and lexical semantics. In *International Conference on Information*, pages 294–302.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,

Anirudh Goyal, Anthony Hartshorn, and Aobo Yang. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wang Guiyuan. 2023. *A Study of Excavated Documents in China*. Routledge, London.

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. 2025. Adapters for altering llm vocabularies: What languages benefit the most? In *International Conference on Representation Learning*, volume 2025, pages 68433–68459.

Xiaowei Han, Lizhen Xu, and Feng Qiao. 2018. Cnn-bilstm-crf model for term extraction in chinese corpus. In *International Conference on Web Information Systems and Applications*, pages 267–274. Springer.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Q Jane. 2014. Ancient times table hidden in chinese bamboo strips.

Zongyuan Jiang, Jiapeng Wang, Jiahuan Cao, Xue Gao, and Lianwen Jin. 2023. Towards better translations from classical to modern chinese: A new dataset and a new method. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 387–399. Springer.

Teruno Kajiura, Shiho Takano, Tatsuya Hiraoka, and Kimio Kuramitsu. 2023. Vocabulary replacement in sentencepiece for domain adaptation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 645–652.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset. *Preprint*, arXiv:2106.01979.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597. Association for Computational Linguistics.

Xiaomeng Pan, Hongfei Wang, Teruaki Oka, and Mamoru Komachi. 2022. Zuo zhuan Ancient Chinese dataset for word sense disambiguation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 129–135, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. 2024. Infonce: Identifying the gap between theory and practice.

Jiu Sha, Mengxiao Zhu, Chong Feng, and Yuming Shang. 2025. Veef-multi-llm: Effective vocabulary expansion and parameter efficient finetuning towards multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7963–7981.

Daqian Shi, Xiaolei Diao, Xu Chen, and Cédric M John. 2025a. Competitive distillation: A simple learning strategy for improving visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2981–2990.

Daqian Shi, Xiaolei Diao, Lida Shi, Hao Tang, Yang Chi, Chuntao Li, and Hao Xu. 2022a. Charformer: A glyph fusion based attentive framework for high-precision character image denoising. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Daqian Shi, Xiaolei Diao, Hao Tang, Xiaomin Li, Hao Xing, and Hao Xu. 2022b. Rcrn: Real-world character image restoration network via skeleton extraction. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Lida Shi, Fausto Giunchiglia, Hongda Zhang, Daqian Shi, Rui Song, Jian Li, Xiaolei Diao, Alan Zhao, and Hao Xu. 2025b. Learn from the best: A universal self-distillation approach with historical logits. *Expert Systems with Applications*, page 129340.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary korean and english. *arXiv preprint arXiv:2205.10019*.

Sho Takase, Ryokan Ri, Shun Kiyono, and Takuya Kato. 2025. Large vocabulary size improves large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1015–1026.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

D Wang and 1 others. 2022. Sikubert and sikuroberta: Construction and application research of pre training models for digital humanities in the complete library

of four branches. In *Lib. Forum*, volume 42, pages 31–43.

Yuting Wei, Yuanxing Xu, Xinru Wei, Yangsimin Yangsimin, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. AC-EVAL: Evaluating Ancient Chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617, Miami, Florida, USA. Association for Computational Linguistics.

Shen Xu, editor. 1981. *Shuowen Jiezi Zhu*. Shanghai Ancient Books Publishing House, Shanghai.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2025. How can we effectively expand the vocabulary of llms with 0.01 gb of target language text? *Computational Linguistics*, pages 1–40.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, and Haoze Sun. 2025a. Baichuan 2: Open large-scale language models. *Preprint*, arXiv:2309.10305.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and Junyang Lin. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Xinyu Yao, Mengdi Wang, Bo Chen, and Xiaobing Zhao. 2025. Wenyangpt: A large language model for classical chinese tasks. *arXiv preprint arXiv:2504.20609*.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, Zheni Zeng, Yuxian Gu, Kun Zhou, and Xuancheng Huan. 2022. Cuge: A chinese language understanding and generation evaluation benchmark. *Preprint*, arXiv:2112.13610.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, and Guanwei Zhang. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Peng Yu and Xin Wang. 2020. Bert-based named entity recognition in chinese twenty-four histories. In *International Conference on Web Information Systems and Applications*, pages 289–301. Springer.

Mingzhe Yue, Daqian Shi, Xiaolei Diao, Shuzhen Guo, Chuntao Li, and Hao Xu. 2025. Ancient character detection based on fine-grained density map. *npj Heritage Science*, 13(1):280.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *CoRR*.

Jundong Zhang, Songhua Yang, Jiangfeng Liu, and 1 others. 2024a. Aigc empowering the revitalization of ancient books on traditional chinese medicine: building the huang-di large language model. *Libr Trib*, 44:103–12.

Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend Ancient Chinese? a preliminary test on ACLUE. In *Proceedings of the Ancient Language Processing Workshop*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yuqing Zhang, Baoyi He, Yihan Chen, Hangqi Li, Han Yue, Shengyu Zhang, Huaiyong Dou, Junchi Yan, Zemin Liu, Yongquan Zhang, and 1 others. 2024b. Philogpt: A philology-oriented large language model for ancient chinese manuscripts with dunhuang as case study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2784–2801.

Zhixiao Zhao, Si Shen, Bin Li, and Xueliang Ma. 2024. Xunziallm.

Zhihan Zhou, Daqian Shi, Rui Song, Lida Shi, Xiaolei Diao, and Hao Xu. 2026. Ancientbench: Towards comprehensive evaluation on excavated and transmitted chinese corpora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 35167–35175.

A The Construction Process of InteChar

Integrated Characters (InteChar) is specifically designed to meet the training requirements of ancient Chinese LLMs, encompassing oracle bone script, traditional Chinese characters, and modern simplified Chinese characters. Data sources include scanned images of oracle bone inscriptions and multiple professional font libraries.

The construction of InteChar follows a four-stage workflow. Firstly, the official Unicode character set is loaded to initialize the character list as a foundational layer, ensuring compatibility with modern computer character systems. Secondly, computer-readable encoded characters from ancient Chinese resources are integrated to enrich the character set. Thirdly, entirely new characters are constructed for glyphs present in the corpus but absent from existing computer character standards. Finally, in collaboration with experts in the field of ancient Chinese, the character list undergoes proofreading and deduplication to guarantee the completeness, accuracy, and uniqueness of each character entry. The resulting integrated characters list thus encompasses both standardized characters and newly encoded glyphs.

We employ the oracle bone script, bronze script, and Warring States script from InteChar as our ancient character vocabulary.

B Ancient Character Feature Module Training Details

We employ contrastive learning to train the ancient character feature module, aiming to position ancient characters as close as possible to embeddings with similar meanings while keeping them as distant as possible from embeddings with dissimilar meanings.

Training Datasets. We constructed glyph embedding training datasets and lexical embedding training datasets respectively. The glyph embedding training set data originates from Interchar (Diao et al., 2026). The data format consists of pairs of ancient characters and their glyphs, where each ancient character corresponds to its component parts. This comprises a total of 13,925 entries, noting that the same ancient character may correspond to different components across various historical periods. The lexical embedding training set data is sourced from ShuoWen (Xu, 1981). The data format consists of pairs of ancient characters and their lexicality, where each ancient character corresponds to its lexical interpretation. This comprises a total of 3,893 entries.

Training Details. We randomly initialize the glyph embedding and lexical embedding, setting ϵ and τ to 1e-12 and 0.1 respectively, with a learning rate of 3e-4 for 12 epochs of training. For both glyph and lexical embeddings, we designated the ancient character and its corresponding glyph/lexicality as positive samples and set the glyph/lexicality of all other ancient characters as negative samples. This approach stems from our intention to train the model's ability to map ancient characters into the modern Chinese representation space, rather than altering the representation space itself. Consequently, even when ancient characters share similar meanings, connections can still be established through implicit associations between ancient characters and modern Chinese characters, and between modern Chinese characters themselves.

C Experimental Implementation Details

We conduct evaluations of ancient character comprehension and representation capabilities. Ancient character comprehension serves to demonstrate the model's proficiency in downstream tasks, while visualization of representation space specifically highlights ACSE's influence on the model's embedding.

Ancient Character Comprehension. We employ AncientBench (Zhou et al., 2026) as our dataset, selecting 4 tasks highly relevant to pre-Qin excavated documents: Radical, Radical Meaning, ExcDoc Word, and Phonetic Loan Character, comprising 8,438, 1,432, 364, and 4,636 pieces of data respectively. During instruction tuning, we first partitioned the samples of each task (four tasks in total) in AncientBench using a 3:7 split (30% for training and 70% for evaluation). We then combined the training portions from all tasks to form a unified instruction-tuning dataset for model fine-tuning, and evaluated the trained model separately on each task's evaluation set. This design approach serves two purposes, on the one hand, it prevents the model from observing validation set answers during training, thereby providing a more direct reflection of its reasoning capabilities. On the other hand, given the strong instruction following ability of large language models, we require only 30% of the data to train the model to fuse features and perform inference. Furthermore, with limited training data available, we aim to utilize as much data as possible for validation, thereby enhancing the persuasiveness of the results.

Visualization of Representation Space. We employ Intechar (Diao et al., 2026) and ShuoWen (Xu, 1981) as data sources to construct a dataset of ancient characters paired with their corresponding glyph/lexicality data, formatted identically to the contrastive learning training set. Cosine similarity was adopted as the metric for relative positioning, as it more accurately reflects the semantic relationships between characters or words. We visualized the positional relationships between ancient characters and their corresponding glyphs/lexicalities by mapping them onto a 2D distribution using multidimensional scaling (MDS).

D Instruction Fine-Tuning Dataset Details

The ancient character vocabulary is a character set (6,498 characters) sourced from Intechar. To introduce prior knowledge about these characters to the LLM, we constructed a glyph dataset (13,925 entries) and a lexical dataset (3,893 entries). The glyph dataset is based on the ancient knowledge graph and contains entries for characters in the

Model	Rad	R.M	Exc	P.L.C
Llama3-8B-Instruct	43.96	38.52	74.52	53.27
ACSE Llama3-8B-Instruct	48.41	47.21	69.80	68.88

Table 4: The ablation study comparing performance with and without ACSE on Llama3-8B-Instruct.

ancient vocabulary together with their component decompositions across different historical periods. The lexical dataset is derived from ShuoWen. We did not use the entire ShuoWen corpus; instead, we extracted the ShuoWen lexicalities that correspond to characters appearing in our ancient vocabulary and constructed the dataset as character–lexicality pairs. Because some characters are undeciphered or lack corresponding explanations in ShuoWen, the lexical dataset contains fewer entries than the full character vocabulary.

E A Discussion of the ACSE Backbone Models

During early experimentation, we evaluated various backbones and selected Qwen. In our evaluations we observed tradeoffs across models: some models that adopt BBPE (e.g., Llama 3 or gpt-2) have limited Chinese proficiency, which harms representation of ancient characters; conversely, some strong Chinese models (e.g., Yi 1.5) do not use BBPE and are therefore not directly compatible with the ACSE token-indexing strategy. In order to further strengthen the argument for the adaptability of our method, we conducted a 5-shot experiment on Llama3-8B-Instructwith the results shown in Table 4.