

InteChar: A Unified Oracle Bone Character List for Ancient Chinese Language Modeling

Xiaolei Diao^{1, 2}, Zhihan Zhou², Lida Shi², Ting Wang³, Ruihua Qi², Hao Xu², Daqian Shi^{1*}

¹Queen Mary University of London ²Jilin University ³Tongji University

Abstract

Constructing historical language models (LMs) plays a crucial role in aiding archaeological provenance studies and understanding ancient cultures. However, existing resources present major challenges for training effective LMs on historical texts. First, the scarcity of historical language samples renders unsupervised learning approaches based on large text corpora highly inefficient, hindering effective pre-training. Moreover, due to the considerable temporal gap and complex evolution of ancient scripts, the absence of comprehensive character encoding schemes limits the digitization and computational processing of ancient texts, particularly in early Chinese writing. To address these challenges, we introduce InteChar, a unified and extensible character list that integrates unencoded oracle bone characters with traditional and modern Chinese. InteChar enables consistent digitization and representation of historical texts, providing a foundation for robust modeling of ancient scripts. To evaluate the effectiveness of InteChar, we construct the Oracle Corpus Set (OracleCS), an ancient Chinese corpus that combines expert-annotated samples with LLM-assisted data augmentation, centered on Chinese oracle bone inscriptions. Extensive experiments show that models trained with InteChar on OracleCS achieve substantial improvements across various historical language understanding tasks, confirming the effectiveness of our approach and establishing a solid foundation for future research in ancient Chinese NLP.

Introduction

Ancient script research has long served as a cornerstone for cultural heritage preservation and the advancement of historical linguistics, enabling scholars to decode lost histories and gain insight into ancient cultures through the inscriptions found on archaeological artifacts. Traditional approaches to this research are largely influenced by human cognitive and learning limitations, which restrict the efficiency of data processing and significantly hinder the decipherment of unknown characters (Diao et al. 2023b). In contrast, recent advances in natural language understanding have demonstrated significant advantages in processing vast and complex corpora, motivating researchers to apply these techniques to large-scale ancient text processing tasks (Tian et al. 2021; Stopponi et al. 2024). One of the key research directions in this field is the development of histor-

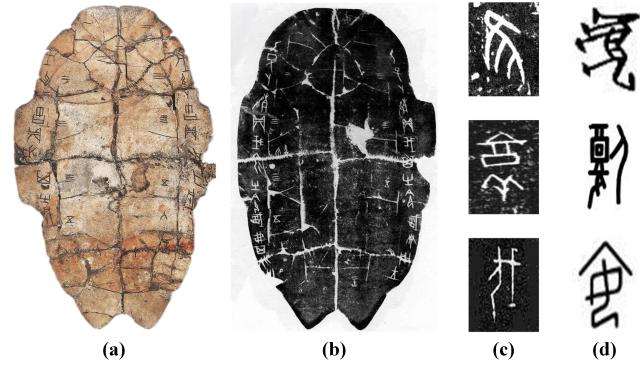


Figure 1: Examples of various types of OBI images. (a) An entire piece of oracle bone. (b) An entire piece of oracle bone rubbing. (c) Oracle character rubbing images. (d) Handwritten oracle character images.

ical language models specifically designed to comprehend ancient textual materials, thereby facilitating tasks such as archaeological inference, historical reconstruction, and cultural analysis (Ross 2023; Koc 2025).

Collecting and constructing appropriate corpora is fundamental to the development of effective historical language models. However, due to the great antiquity of ancient scripts and their complex evolution over time, excavated documents represented by Oracle Bone Inscriptions (OBI) frequently contain a large number of unencoded characters, which poses substantial challenges for digitization. As illustrated in Figure 1, the common approach currently involves storing these characters as images (Shi et al. 2022b; Guan et al. 2024; Wang et al. 2024; Gao et al. 2024), which are often in handwritten or rubbing form, rather than as machine-encoded text. Furthermore, the limited number and preservation issues of excavated artifacts result in relatively scarce corpus samples for these ancient languages (Chi et al. 2022; Diao et al. 2023a). For example, only about 5,000 complete oracle bone pieces have been unearthed, yielding merely 15,000 sentences that contain more than five characters. Consequently, the combination of vast numbers of unencoded characters and the scarcity of sentence-level samples poses a significant challenge to the construction of an effective and usable corpus for ancient texts.

*Corresponding Author.

Research on ancient language models remains at an early stage. Some studies (Guo et al. 2023; Wang et al. 2023; Liu et al. 2024) have attempted to repurpose modern character encoding schemes to represent ancient scripts, training word embeddings and constructing corpora based on these representations. However, such approaches typically cover only a subset of high-frequency ancient characters, while excluding a large number of low-frequency or unencoded characters. This exclusion is not trivial: in the context of ancient texts, where language resources are extremely limited and the writing system often highly contextual, even a single low-frequency character may carry unique semantic, historical, or cultural significance (Diao et al. 2023b). Each character, regardless of its frequency, can be crucial for accurately reconstructing meaning, understanding rare expressions, or tracing cultural and linguistic developments. Therefore, models that omit these characters suffer from incomplete representations and risk significant information loss during training (Zhang and Li 2023). Other efforts have explored directly training language models on limited transcriptions from excavated artifacts (Chi et al. 2022), but these are similarly constrained by the scarcity of annotated data, which makes unsupervised pre-training inefficient and affects effective semantic learning. The extremely low frequency of many ancient characters further complicates this challenge, as their meanings cannot be reliably inferred from the surrounding context alone.

To address the above challenges, we propose a novel corpus to support historical language modeling. A key component of our approach is InteChar, a unified and extensible character set that incorporates oracle bone characters not covered by existing encoding standards. We introduce a standardized digitization pipeline that converts scanned images containing OBIs into machine-readable text, encoded in a format compatible with modern character standards. InteChar enables a consistent and comprehensive digital representation of ancient scripts, providing a solid foundation for subsequent corpus construction and model training. Furthermore, although many ancient characters appear infrequently, their latent mappings to modern Chinese can be reinforced using data distillation techniques within a pre-training paradigm. By augmenting the corpus with enhanced samples and integrating specialized pre-training strategies, models can acquire semantic representations of these rare characters more rapidly. Following this strategy, we construct OracleCS, a corpus of excavated transcriptions of oracle bone inscriptions designed to support the training of historical Chinese language models in low-resource settings. In addition, we construct a multi-task benchmark to quantitatively evaluate the performance of language models on understanding ancient scripts. To our knowledge, this work is the first to systematically incorporate excavated oracle characters, including undeciphered ones, into LM’s evaluation pipelines. The main contributions include:

- We construct InteChar, a unified and extensible Unicode-compatible character list that integrates previously unencoded oracle bone characters alongside traditional and modern Chinese characters, enabling consistent and comprehensive digitization of ancient texts.

- Based on InteChar, we build OracleCS, a corpus that prominently features oracle bone transcriptions, and develop a benchmark to systematically evaluate language models on ancient Chinese language understanding.
- Extensive experiments demonstrate that models trained on OracleCS with InteChar significantly outperform baselines across both embedding-based evaluations and downstream fine-tuning tasks.

Related Work

Ancient Character Tasks

In recent years, the study of ancient Chinese characters has gained increasing attention due to its value in historical linguistics, archaeology, and cultural heritage. Advances in deep learning have greatly supported this field, especially in tasks like character recognition (Lin et al. 2022), detection (Yue et al. 2025), and restoration (Shi et al. 2022a; Li et al. 2023). Early studies primarily focused on image-based recognition tasks, such as oracle bone inscription classification using computer vision techniques. For example, Zhang et al. (Zhang et al. 2021) employed a Siamese network to match rubbing images with template images from oracle character databases. Later studies have expanded beyond recognition to include glyph identification and structure analysis. RZCR (Diao et al. 2023b) combines radical and structure features for character recognition, while LUC (Gao et al. 2024) added radical and domain-specific features to improve character retrieval. (Chi et al. 2022) propose an ancient Chinese knowledge graph ZiNet that links glyphs and radicals across time. Generative methods have also been explored. (Guan et al. 2024) proposed a diffusion-based approach to generate possible modern forms from ancient characters, helping with transcription and understanding how characters evolved. Despite these advances, such studies only focus on deciphered characters. The interpretation of entirely unknown characters remains an open and challenging problem.

Ancient Chinese Corpus Collection

A key challenge in using NLP for ancient languages is the lack of annotated data. Ancient Chinese texts, e.g., oracle bones and bronze inscriptions, are typically available only as noisy, fragmented images with limited annotations (Shi et al. 2022b; Diao et al. 2025). This low-resource scenario necessitates innovative data collection and augmentation strategies. Recent work on word segmentation has explored using language models to create synthetic training data. For example, (Shen et al. 2022) used an LSTM model (Hochreiter and Schmidhuber 1997) to generate labeled samples, and (Feng and Li 2023) applied distant supervision with parallel texts to build augmented datasets. Expert-annotated corpora are also essential. For instance, the CHisIEC dataset (Tang et al. 2024) combines expert knowledge with text analysis to extract relations from historical texts.

Benchmarks of Ancient Language Models

Evaluating the performance of language models trained on ancient scripts poses unique challenges. Recent research has

focused on building benchmark datasets to assess model capabilities on ancient Chinese texts. For example, FSPC (Shao et al. 2021) and CCMP (Li et al. 2021) provide corpora for classical poetry comprehension tasks, while CUGE¹ (Yao et al. 2021) extends CCMP by adding a poetry matching subtask. Zinin and Xu (Zinin and Xu 2020) compiled historical travel texts and other ancient corpora to enrich data diversity for downstream tasks. Other studies have introduced tasks such as syntactic analysis, topic mining, and sentiment classification (Pan et al. 2022; Wang and Ren 2022; Liu et al. 2022). AC-EVAL (Wei et al. 2024) integrates multiple datasets and tasks into a unified evaluation suite for ancient Chinese understanding. WenMind (Cao et al. 2024a) focuses on Chinese classical literature and language arts, containing 4,875 question–answer pairs across 42 fine-grained tasks, offering a comprehensive benchmark for evaluating LLMs in this domain. Fùxì (Zhao et al. 2025) introduces a benchmark covering 21 tasks aimed at both understanding and generation, including novel tasks such as poetry composition and couplet completion, making it particularly suited for generation-oriented ancient Chinese tasks.

In summary, although significant progress has been made in developing models for the recognition and interpretation of ancient Chinese characters, key challenges remain. The most crucial problem is that all existing benchmarks and corpora only cover encoded texts, ignoring characters from unearthed artifacts that have not been standardized. This greatly limits the ability of language models to learn from ancient texts and manuscripts that contain rich historical and cultural information.

The Proposed Method

This section introduces the construction of two key resources: the InteChar Unicode character list and the OracleCS corpus. InteChar provides a unified encoding for ancient characters, while OracleCS offers a pretraining dataset for oracle bone script. Both of them form the foundation for downstream modeling and evaluation.

InteChar Character list Construction

Research on ancient scripts underscores the importance of building a comprehensive Unicode character list to support the development of historical language models. In this work, we construct a unified and structured character set named Integrated Characters (InteChar), which includes oracle bone characters, traditional Chinese characters, and modern simplified characters. InteChar is specifically designed to meet the needs of training models on ancient Chinese texts. We integrate multiple data sources, including modern Unicode character sets, scanned images of oracle bone inscriptions, and specialized font libraries, to produce a complete and standardized character list.

The construction of InteChar follows a four-stage workflow aimed at building a unified and extensible character inventory for ancient Chinese scripts. First, we initialize the character list by loading the official Unicode charac-

ter set², which serves as the foundational layer for compatibility with modern natural language processing systems. Second, we enrich the list by incorporating encoded characters from widely adopted machine-readable ancient Chinese resources, selecting only those characters that also appear in our curated corpus to ensure relevance and avoid redundancy. Third, we construct entirely new characters for glyphs present in the corpus but absent from existing standards. Finally, we conduct expert-guided proofreading and de-duplication to ensure the integrity, accuracy, and uniqueness of each character entry. The resulting InteChar character set contains both standardized and newly encoded characters, offering robust support for historical language modeling and digital processing of ancient texts.

Initial Character List Construction. The initial character list is constructed by loading the official Unicode character set, which defines standardized code points for characters from virtually all major writing systems worldwide. Among them, Unicode includes more than 90,000 encoded CJK characters³. This serves as the foundational encoding standard for modern natural language processing, ensuring consistency across platforms and compatibility with mainstream language models.

Integration of Existing Encoded Characters. To fully leverage existing resources and reduce manual overhead, we integrate previously encoded ancient characters from widely adopted machine-readable libraries. A primary source is the Zhongjian Library collection published by Zhonghua Book Company⁴, a commonly used resource among paleographers that includes 16 historical font sets initially. These fonts include well-attested glyphs from oracle bones, bronze inscriptions, bamboo manuscripts, and other early Chinese scripts. Our focus is on the oracle bone subset of the Zhongjian Library. To ensure relevance and avoid redundancy, we apply a strict filtering strategy: only characters that appear both in the Zhongjian Library and in our curated corpus are retained. This ensures that all included characters are not only well-formed but also actively used in authentic historical contexts. When a character appears in multiple font sets, we preserve only one representative form to avoid duplication. For each retained glyph, we extract its graphical representation and record the associated metadata in InteChar, including font source and an internally assigned code point within Zhongjian Library. This integration stage enables the reuse of trusted typographic resources and provides a cost-effective foundation for character set expansion.

Construction of New Characters. While the integration of existing resources significantly reduces manual effort, a large number of characters in our corpus remain unencoded in both Unicode and historical font libraries, especially from excavated oracle inscriptions. These characters often correspond to undeciphered or low-frequency glyphs that nonetheless carry important contextual and linguistic value. To address this gap, we construct entirely new characters using a semi-automated pipeline that combines com-

²<https://en.wikipedia.org/wiki/Unicode>

³https://en.wikipedia.org/wiki/CJK_Unified_Ideographs

⁴<https://www.ancientbooks.cn/helpcore?font>

¹<https://cuge.baai.ac.cn>.

Oracle Bone Characters										
Glyphs	數	鼓	宜	耘	僕					
Unicodes	F0236	F0237	F023A	F023B	F0249					

(a)



(b)

Figure 2: Examples of Unicode character lists. (a). Oracle bone Character images with their corresponding standardized glyphs and Unicode in InteChar. (b). Examples of TrueType Font in InteChar.

puter vision techniques with expert validation. Following previous studies, we rely on three key sources to identify candidate characters for construction: oracle bone images, domain-specific corpora, and existing font resources. Characters that appear in our training corpus but lack a corresponding encoding in Unicode or the Zhongjian Library are flagged as candidates for reconstruction.

We observe that while many oracle glyphs are complex and unique, a large number of their subcomponents, namely radicals, have been extensively studied and can be linked to known components in traditional Chinese characters. Instead of requiring experts to manually trace entire glyphs stroke by stroke, we adopt a radical-based recognition method that detects familiar structural units to enable compositional reconstruction. This approach significantly improves both the efficiency and scalability of new character creation. To efficiently construct new character entries without requiring experts to manually trace every stroke, we apply a radical recognition method that automatically identifies radicals within complex glyphs. By detecting known radicals, we can reconstruct new characters radical by radical, rather than stroke by stroke, significantly accelerating the expansion of the character list. In our implementation, we adopt the radical recognition method proposed by (Diao et al. 2023a) to output candidate radicals, which is based on an object detection model trained to identify oracle radicals within noisy, real-world character images. The overall pipeline for new character construction consists of the following seven steps:

- **Image Collection and Preprocessing:** Oracle bone inscription images are collected from archaeological publications and processed via resizing, contrast normalization, and geometric alignment (e.g., vertical flipping).
- **Radical Recognition:** The preprocessed images are passed through the radical recognition model to predict the categories of radicals within each glyph.
- **Standardization of Components:** Mapping predicted radicals to modern Chinese equivalents where applicable, forming an intermediate compositional glyphs.
- **Expert Verification:** Paleographers manually verify the radical composition, correct misclassifications, and make final glyph adjustments based on domain knowledge.

- **Vectorization:** Validated glyphs are redrawn as scalable vector graphics, conforming to a consistent visual style aligned with InteChar’s typographic standard.
- **Code Point Assignment:** Each reconstructed glyph is assigned a new internal code point using a Unicode-style format that supports future interoperability and systematic expansion.
- **Character Integration:** The finalized character is incorporated into InteChar.

This pipeline allows us to systematically digitize and encode characters that were previously inaccessible to computational models. As a result, oracle bone characters with no prior encoding can now be consistently represented, searched, and used in downstream language modeling tasks. Figure 2(a) illustrates examples of newly constructed characters, showing the original image, reconstructed glyph, and assigned code point within InteChar.

Expert-Guided Proofreading. When finished with the construction of the InteChar, we invite domain experts in paleography to validate InteChar through human-in-the-loop proofreading. Applying Siamese networks (Melekhov, Kannala, and Rahtu 2016), we compute glyph similarities to identify potential duplicates and present them with corresponding encodings for expert review. The InteChar we built contained a total of 11,288 characters. Figure 2(b) presents the TrueType Font of InteChar, namely the “.ttf” file. Importantly, InteChar is designed to be continuously updatable. Any new additions to the list follow the same construction pipeline. This process addresses the long-standing issues of incomplete digitization and data sparsity in ancient Chinese texts, and enables robust training and analysis of early scripts such as oracle bones.

OracleCS Corpus Construction

To support embedding-level representation learning and downstream fine-tuning for ancient Chinese language understanding, we construct the Oracle Corpus Set (OracleCS), a linguistically curated corpus specifically focused on oracle bone inscriptions and ancient Chinese texts. It serves as a foundational resource for evaluating and adapting modern language models to low-resource historical scripts.

OracleCS is constructed under a standardized pipeline that integrates expert curation with automated data enrichment. The process begins with domain experts in paleography and historical Chinese linguistics manually selecting and annotating high-quality samples from archaeological literature and oracle bone rubbings⁵. These samples include both deciphered and undeciphered oracle characters encoded in **InteChar**, and form the core of the corpus. In addition to the main textual data, OracleCS includes glyph-level and semantic annotations for individual characters. These include radical decompositions and definitions extracted from classical lexicons such as Shuowenjiezi⁶ and The Great Chinese Dictionary⁷. For characters not yet deciphered, semantic annotations are left blank. The corpus also incorporates a selection of pre-Qin classics⁸, including Analects, Spring and Autumn Annals, Mencius, Xunzi, etc.

To address the scarcity of annotated ancient texts, we adopt data augmentation strategies to further enrich OracleCS. Specifically, we introduce instruction-tuning samples that combine task descriptions with input-output demonstrations. These instruction-following examples simulate realistic usage scenarios and cover a wide range of sentence-level and character-level tasks, including sentence translation, synonym substitution, glyph structure analysis, character decomposition, and semantic prediction. By integrating explicit instructions with concrete demonstrations, the corpus is expanded in scale and enhanced in task diversity and linguistic granularity. This enables the model to more effectively learn both high-level semantic mappings and low-level structural associations, improving its ability to evaluate the complexities of ancient text processing.

Based on these datasets, we construct a benchmark for evaluating historical language understanding. Our framework supports two complementary modes of evaluation, including embedding evaluation tasks and downstream fine-tuning tasks. Notably, this work is the first to incorporate excavated oracle characters into systematic evaluation pipelines for large language models, including oracle bone characters that remain undeciphered.

Experiments and Discussions

This section presents the experimental evaluation of InteChar and OracleCS. We detail the datasets used, the experimental setup, baseline comparisons, and downstream task evaluations to demonstrate the effectiveness of our methodology. In this section, we present a comprehensive evaluation of our proposed OracleCS through two sets of experiments. The first set assesses the embedding capability of language models pre-trained with the extended oracle vocabulary, while the second set evaluates downstream performance on several fine-tuning tasks. In all experiments, we trained each baseline model both with and without the addition of the InteChar character list, and compared their performance on the same tasks.

⁵<https://jgw.aynu.edu.cn/home/index.html>

⁶<https://www.shuowen.cn/>

⁷<https://www.hanyudacidian.cn/>

⁸<https://ctext.org/pre-qin-and-han/zhs>

Experimental Setup

Datasets. Our experiments are conducted based on the proposed OracleCS dataset, which comprises both excavated texts and classical Chinese literature. The dataset contains approximately 11,288 unique Chinese characters and a total of 173,459 annotated samples. Each sample includes radical decomposition information, and, where applicable, a mapping to its corresponding modern Chinese word is also provided. This structured annotation enables more accurate learning of character semantics and facilitates model understanding of the intricate relationship between ancient and modern language forms.

Baselines. All evaluations are conducted on ten models. We select three classic baselines, including BERT (Devlin et al. 2019), Llama-3-8B (Touvron et al. 2023), and GPT-2 (Brown et al. 2020), three language models designed for Chinese, including MiniRBT (Cui et al. 2021), guwenBERT-base⁹ (Wang et al. 2023), and sikuBERT (Liu et al. 2024), two state-of-the-art LLMs, including Qwen-7B-Chat (Bai et al. 2023) and GLM-4-9B (GLM et al. 2024), and two state-of-the-art LLMs designed for ancient Chinese, including XunziALLM¹⁰ and TongGu-LLM (Cao et al. 2024b).

Implementation details. The experiments are run on a high-performance server equipped with eight HUAWEI Ascend-D910b NPU under an Ubuntu-based environment. We implement our models with PyTorch and set unified hyperparameters for every models. For embedding evaluation, models are trained for 10 epochs with a batch size of 32 and an initial learning rate of 3e-5. Optimization is performed using AdamW, and early stopping is applied based on NDCG@10 on the development set. For fine-tuning evaluation, we set the batch size to 32 and the learning rate to 1e-5. It is performed for 10 epochs using the AdamW optimizer and CrossEntropy loss. The model checkpoint with the highest validation score is selected for final evaluation.

Embedding Evaluation

To assess the semantic representation capabilities of pre-trained models under different character list settings, we design two embedding-based evaluation tasks: (1) Cloze Completion on Oracle Bone Inscriptions, and (2) Commentary-to-Text Retrieval on Canonical Texts. These tasks evaluate how well the character-level embeddings capture contextual and semantic information in a zero-shot setting, without task-specific fine-tuning. For each model, we replace the original character embedding layer with a newly initialized embedding matrix based on either the original character list or the proposed InteChar, while keeping the pretrained model backbone frozen. These embeddings are trained on the OracleCS corpus to adapt to ancient character representations. During evaluation, we extract final-layer embeddings and compute similarity scores between masked inputs and candidates or between paired sentence-level inputs. Performance is reported using standard ranking metrics: Nor-

⁹<https://github.com/ethan-yt/guwenbert>.

¹⁰<https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM>.

Models	NDCG@10		MRR@10		NDCG@20		MRR@20	
	Origin	InteChar	Origin	InteChar	Origin	InteChar	Origin	InteChar
BERT (Devlin et al. 2019)	0.167	0.515	0.134	0.375	0.163	0.453	0.127	0.312
Llama-3-8B (Touvron et al. 2023)	0.172	0.518	0.143	0.463	0.212	0.546	0.115	0.334
GPT-2 (Brown et al. 2020)	0.216	0.584	0.168	0.534	0.224	0.643	0.176	0.488
MiniRBT (Cui et al. 2021)	0.184	0.538	0.138	0.413	0.154	0.441	0.121	0.358
guwenBERT-base ⁹ (Wang et al. 2023)	0.204	0.565	0.156	0.526	0.168	0.480	0.132	0.386
sikuBERT (Liu et al. 2024)	0.195	0.553	0.163	0.488	0.182	0.513	0.143	0.423
Qwen-7B-Chat (Bai et al. 2023)	0.302	0.842	0.254	0.736	0.280	0.795	0.228	0.639
GLM-4-9B (GLM et al. 2024)	0.274	0.808	0.278	0.752	0.266	0.762	0.209	0.618
XunziALLM ¹⁰	0.261	0.765	0.225	0.675	0.252	0.725	0.232	0.651
TongGu-LLM (Cao et al. 2024b)	0.238	0.723	0.213	0.638	0.238	0.683	0.187	0.553

Table 1: Embedding-based evaluation results on the Cloze task using excavated oracle texts. Models use frozen backbones and are equipped with newly trained embedding layers based on either the original character list or the InteChar character list. Performance is measured by NDCG@ k and MRR@ k , where $k = 10$ or 20 .

Models	NDCG@400		MRR@400		NDCG@500		MRR@500	
	Origin	InteChar	Origin	InteChar	Origin	InteChar	Origin	InteChar
BERT (Devlin et al. 2019)	0.346	0.376	0.136	0.207	0.327	0.452	0.112	0.182
Llama-3-8B (Touvron et al. 2023)	0.298	0.325	0.098	0.152	0.281	0.388	0.085	0.137
GPT-2 (Brown et al. 2020)	0.320	0.347	0.115	0.175	0.300	0.415	0.095	0.155
MiniRBT (Cui et al. 2021)	0.358	0.390	0.145	0.221	0.340	0.471	0.120	0.195
guwenBERT-base ⁹ (Wang et al. 2023)	0.375	0.406	0.155	0.237	0.355	0.492	0.128	0.208
sikuBERT (Liu et al. 2024)	0.385	0.417	0.162	0.247	0.365	0.506	0.133	0.216
Qwen-7B-Chat (Bai et al. 2023)	0.435	0.472	0.202	0.308	0.418	0.579	0.168	0.272
GLM-4-9B (GLM et al. 2024)	0.448	0.486	0.211	0.322	0.432	0.595	0.176	0.285
XunziALLM ¹⁰	0.402	0.436	0.182	0.278	0.394	0.540	0.148	0.240
TongGu-LLM (Cao et al. 2024b)	0.390	0.424	0.175	0.267	0.378	0.523	0.141	0.229

Table 2: Embedding-based evaluation results on the Commentary-to-Text Retrieval task. Models use frozen backbones and are equipped with newly trained embedding layers based on either the original character list or the InteChar character list. Performance is evaluated with NDCG@ k and MRR@ k , where $k = 400$ or 500 .

malized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR).

Cloze Completion on Oracle Bone Inscriptions. This task focuses exclusively on oracle bone inscriptions from excavated sources, aiming to test the word embedding ability of models to semantically distinguish oracle characters in context. Given a sentence with one character masked, the model is asked to select the correct character from a limited set of candidates based on embedding similarity. Each candidate set is predefined (e.g., @ k indicates k options per instance). This test set consists of 15,416 cloze instances, with 12,416 instances for training and 3,000 for evaluation. Each instance includes a masked sentence and a set of candidate characters (including one ground truth).

Table 1 presents the performance of 10 representative models across four ranking metrics: NDCG@10, MRR@10, NDCG@20, and MRR@20. Across all metrics, models using InteChar consistently outperform their original counterparts. For example, the MRR@10 of GPT improves from 0.168 to 0.534, and BERT from 0.134 to 0.375, demonstrating better top-ranked prediction accuracy. Larger mod-

els show even greater gains: Qwen2.5-Omni-7B improves from 0.302 to 0.842 in NDCG@10, and from 0.254 to 0.736 in MRR@10. These results validate the effectiveness of InteChar in enhancing representation learning, particularly for low-resource ancient scripts. The enriched character semantics help models capture context more effectively.

Commentary-to-Text Retrieval on Canonical Texts. This task evaluates sentence-level semantic alignment between modern commentaries and classical Chinese texts, aiming to assess sentence-level semantic understanding and retrieval capacity. Given a modern commentary, the model retrieves the corresponding original sentence from a large candidate pool based on sentence embedding similarity. The test set contains 896 commentary queries and 12,141 classical text candidates. Each model computes sentence-level embeddings for both, ranked by similarity, and evaluated using standard retrieval metrics such as NDCG@ k and MRR@ k .

As shown in Table 2, InteChar-enhanced models again yield notable improvements. For instance, when $k = 500$, GLM-4-9B improves its MRR@500 from 0.176 to 0.285, and Qwen2.5-Omni-7B increases its NDCG@500 from

Models	Translation		Polysemous Matching		Word Parsing		Average	
	origin	InteChar	origin	InteChar	origin	InteChar	origin	InteChar
BERT (Devlin et al. 2019)	92.75	93.01	86.69	87.07	90.54	91.59	89.99	90.56
Llama-3-8B (Touvron et al. 2023)	83.52	83.43	80.28	80.51	80.43	80.89	81.41	81.61
GPT-2 (Brown et al. 2020)	90.27	90.84	86.86	87.23	88.67	89.27	88.60	89.11
MiniRBT (Cui et al. 2021)	91.34	91.69	86.65	87.12	89.32	89.86	89.10	89.56
guwenBERT-base ⁹ (Wang et al. 2023)	92.86	93.50	87.73	88.48	91.26	91.88	90.62	91.29
sikuBERT (Liu et al. 2024)	93.21	93.88	86.48	87.52	90.84	91.32	90.18	90.91
Qwen-7B-Chat (Bai et al. 2023)	94.37	95.06	89.23	90.16	93.36	93.96	92.32	93.06
GLM-4-9B (GLM et al. 2024)	92.98	93.35	87.72	88.34	94.27	94.79	91.66	92.16
XunziALLM ¹⁰	93.53	94.31	91.51	92.23	92.78	93.28	92.61	93.27
TongGu-LLM (Cao et al. 2024b)	94.12	94.84	90.45	91.27	92.06	92.65	92.21	92.92

Table 3: Fine-tuning results (%) on three downstream tasks, including Ancient Chinese Translation, Polysemous Word Matching, and Word Parsing. Each model is adapted on the OracleCS using either the original character list or our proposed InteChar character list. The Average column represents the average accuracy across all three tasks.

0.418 to 0.579. These gains reflect the enhanced capacity of InteChar to bridge semantic gaps between classical and modern Chinese expressions.

Overall, these two tasks evaluate both fine-grained (character-level) and coarse-grained (sentence-level) semantic capabilities. The consistent performance improvements demonstrate that InteChar facilitates more robust and discriminative character embeddings, enabling better semantic matching in low-resource, zero-shot scenarios.

Fine-tuning Evaluation

In addition to embedding-based evaluation, we further validate the effectiveness of our proposed InteChar character list under fine-tuning settings. Specifically, we consider three downstream tasks: Ancient Chinese Translation, Polysemous Word Matching, and Word Parsing. Details include:

- Ancient Chinese Translation is a sentence-level task that requires the model to align ancient Chinese texts with their modern Chinese counterparts.
- Polysemous Word Matching is a binary classification task where the model is given a sentence and asked to determine whether a specified character in context matches a given semantic interpretation.
- Word Parsing is a character-level task where the model selects the appropriate interpretation of an ancient character based on learned semantics.

All experiments are conducted on annotated subsets of the OracleCS dataset, using either the original character list or InteChar, with separate training and test sets for each task. The Ancient Chinese Translation task includes 15,868 training and 10,578 test samples; Polysemous Word Matching has 33,380 training and 22,253 test samples; and Word Parsing consists of 81,929 training and 54,619 test samples. These tasks cover different levels of linguistic granularity, from sentence-level translation to character-level parsing, and together form a comprehensive benchmark for evaluating historical language understanding.

We adopt parameter-efficient fine-tuning using LoRA (Hu et al. 2022) to adapt each model to downstream tasks. In

all cases, we freeze the pretrained model backbone and update only the low-rank adaptation layers and task-specific output heads. As shown in Table 3, the results demonstrate that training with InteChar consistently improves performance across all tasks and models. Compared to the original character list, models with InteChar achieve better generalization and semantic alignment. For example, TongGu-LLM reaches 94.84 on translation and 92.65 on parsing, yielding an overall accuracy of 92.92. The average accuracy across all tasks also improves for every model. For instance, Qwen2.5-Omni-7B rises from 92.32 to 93.06. These improvements suggest that InteChar enhances both shallow and deep linguistic modeling. Compared to embedding-based evaluations, the fine-tuning experiments provide complementary evidence. While embedding evaluation focuses on the quality of newly trained character embeddings under a frozen backbone, fine-tuning further adapts models through parameter-efficient tuning for specific tasks. The consistent gains across both settings confirm that InteChar significantly improves language modeling for ancient texts.

Conclusion

This paper focuses on addressing key challenges in training language models for historical Chinese texts, including the poor performance of conventional language models on sparse ancient data and the lack of unified digital representations for ancient characters. One of our contributions is the construction of InteChar, a unified and extensible character set that incorporates unencoded oracle characters alongside traditional and modern Chinese, enabling consistent representation across scanned images, font libraries, and annotated corpora. We further integrate expert-curated samples with LLM-assisted data augmentation to construct a high-quality training corpus, OracleCS, and evaluate models on both cloze-style completion for excavated texts and commentary-to-text retrieval on classical literature. Experimental results show that models equipped with InteChar significantly outperform those using the original character list in both embedding-based and fine-tuning tasks, particularly in handling rare or unencoded characters.

References

- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. arXiv:2309.16609.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Cao, J.; Liu, Y.; Shi, Y.; Ding, K.; and Jin, L. 2024a. WenMind: A comprehensive benchmark for evaluating large language models in Chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 37: 51358–51410.
- Cao, J.; Peng, D.; Zhang, P.; Shi, Y.; Liu, Y.; Ding, K.; and Jin, L. 2024b. TongGu: Mastering Classical Chinese Understanding with Knowledge-Grounded Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4196–4210.
- Chi, Y.; Giunchiglia, F.; Shi, D.; Diao, X.; Li, C.; and Xu, H. 2022. ZiNet: Linking Chinese Characters Spanning Three Thousand Years. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3061–3070.
- Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.
- Diao, X.; Shi, D.; Cao, W.; Wang, T.; Qi, R.; Li, C.; and Xu, H. 2025. Oracle bone inscription image restoration via glyph extraction. *npj Heritage Science*, 13(1): 321.
- Diao, X.; Shi, D.; Li, J.; Shi, L.; Yue, M.; Qi, R.; Li, C.; and Xu, H. 2023a. Toward Zero-shot Character Recognition: A Gold Standard Dataset with Radical-level Annotations. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6869–6877.
- Diao, X.; Shi, D.; Tang, H.; Shen, Q.; Li, Y.; Wu, L.; and Xu, H. 2023b. RZCR: Zero-shot Character Recognition via Radical-based Reasoning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*.
- Feng, S.; and Li, P. 2023. Ancient Chinese word segmentation and part-of-speech tagging using distant supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Gao, F.; Chen, X.; Li, B.; Liu, Y.; Jiang, R.; and Han, Y. 2024. Linking unknown characters via oracle bone inscriptions retrieval. *Multimedia Systems*, 30(3): 125.
- GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Guan, H.; Yang, H.; Wang, X.; Han, S.; Liu, Y.; Jin, L.; Bai, X.; and Liu, Y. 2024. Deciphering Oracle Bone Language with Diffusion Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15554–15567.
- Guo, G.; Yang, J.; Lu, F.; Qin, J.; Tang, T.; and Zhao, W. X. 2023. Towards effective ancient chinese translation: Dataset, model, and evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 416–427. Springer.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Koc, V. 2025. Exploring the Role of Language Models in Deciphering and Preserving Ancient Languages. *Asian American Research Letters Journal*, 2(1): 75–82.
- Li, J.; Wang, Q.-F.; Huang, K.; Yang, X.; Zhang, R.; and Goulermas, J. Y. 2023. Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognition*, 140: 109534.
- Li, W.; Qi, F.; Sun, M.; Yi, X.; and Zhang, J. 2021. Ccpm: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.
- Lin, X.; Chen, S.; Zhao, F.; and Qiu, X. 2022. Radical-based extract and recognition networks for Oracle character recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 25(3): 219–235.
- Liu, C.; Wang, D.; Zhao, Z.; Hu, D.; Wu, M.; Lin, L.; Liu, J.; Zhang, H.; Shen, S.; Li, B.; et al. 2024. SikuGPT: A Generative Pre-trained Model for Intelligent Information Processing of Ancient Texts from the Perspective of Digital Humanities. *ACM Journal on Computing and Cultural Heritage*, 17(4): 1–17.
- Liu, M.; Xiang, J.; Xia, X.; and Hu, H. 2022. Contrastive Learning between Classical and Modern Chinese for Classical Chinese Machine Reading Comprehension. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).
- Melekhov, I.; Kannala, J.; and Rahtu, E. 2016. Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, 378–383. IEEE.
- Pan, X.; Wang, H.; Oka, T.; and Komachi, M. 2022. Zuo Zhuan Ancient Chinese Dataset for Word Sense Disambiguation. In Ippolito, D.; Li, L. H.; Pacheco, M. L.; Chen, D.; and Xue, N., eds., *Proceedings of the 2022 Conference*

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 129–135. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics.
- Ross, E. A. 2023. A new frontier: AI and ancient language pedagogy. *Journal of Classics Teaching*, 24(48): 143–161.
- Shao, Y.; Shao, T.; Wang, M.; Wang, P.; and Gao, J. 2021. A Sentiment and Style Controllable Approach for Chinese Poetry Generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM ’21, 4784–4788. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.
- Shen, Y.; Li, J.; Huang, S.; Zhou, Y.; Xie, X.; and Zhao, Q. 2022. Data augmentation for low-resource word segmentation and pos tagging of ancient chinese texts. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 169–173.
- Shi, D.; Diao, X.; Shi, L.; Tang, H.; Chi, Y.; Li, C.; and Xu, H. 2022a. CharFormer: A Glyph Fusion based Attentive Framework for High-precision Character Image Denoising. In *Proceedings of the 30th ACM International Conference on Multimedia*.
- Shi, D.; Diao, X.; Tang, H.; Li, X.; Xing, H.; and Xu, H. 2022b. RCRN: Real-world Character Image Restoration Network via Skeleton Extraction. In *Proceedings of the 30th ACM International Conference on Multimedia*.
- Stopponi, S.; Pedrazzini, N.; Peels-Matthey, S.; McGillivray, B.; and Nissim, M. 2024. Natural Language Processing for Ancient Greek: Design, advantages and challenges of language models. *Diachronica*, 41(3): 414–435.
- Tang, X.; Su, Q.; Wang, J.; and Deng, Z. 2024. CHisIEC: An Information Extraction Corpus for Ancient Chinese History. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3192–3202. Torino, Italia: ELRA and ICCL.
- Tian, H.; Yang, K.; Liu, D.; and Lv, J. 2021. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Wang, D.; Liu, C.; Zhao, Z.; Shen, S.; Liu, L.; Li, B.; Hu, H.; Wu, M.; Lin, L.; Zhao, X.; et al. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.
- Wang, P.; and Ren, Z. 2022. The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS. In Sprugnoli, R.; and Passarotti, M., eds., *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 164–168. Marseille, France: European Language Resources Association.
- Wang, P.; Zhang, K.; Wang, X.; Han, S.; Liu, Y.; Wan, J.; Guan, H.; Kuang, Z.; Jin, L.; Bai, X.; et al. 2024. An open dataset for oracle bone character recognition and decipherment. *Scientific Data*, 11(1): 976.
- Wei, Y.; Xu, Y.; Wei, X.; Yangsimin, Y.; Zhu, Y.; Li, Y.; Liu, D.; and Wu, B. 2024. AC-EVAL: Evaluating Ancient Chinese Language Understanding in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1600–1617.
- Yao, Y.; Dong, Q.; Guan, J.; Cao, B.; Zhang, Z.; Xiao, C.; Wang, X.; Qi, F.; Bao, J.; Nie, J.; et al. 2021. Cuge: A chinese language understanding and generation evaluation benchmark. *arXiv preprint arXiv:2112.13610*.
- Yue, M.; Shi, D.; Diao, X.; Guo, S.; Li, C.; and Xu, H. 2025. Ancient character detection based on fine-grained density map. *npj Heritage Science*, 13(1): 280.
- Zhang, C.; Zong, R.; Cao, S.; Men, Y.; and Mo, B. 2021. AI-powered oracle bone inscriptions recognition and fragments rejoining. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 5309–5311.
- Zhang, Y.; and Li, H. 2023. Can large language model comprehend ancient chinese? a preliminary test on acle. In *Proceedings of the Recent Advances in Natural Language Processing: RANLP 2023*.
- Zhao, S.; Zhou, Y.; Ren, Y.; Chen, Z.; Jia, C.; Zhe, F.; Long, Z.; Liu, S.; and Lan, M. 2025. F\ux\i: A Benchmark for Evaluating Language Models on Ancient Chinese Text Understanding and Generation. *arXiv preprint arXiv:2503.15837*.
- Zinin, S.; and Xu, Y. 2020. Corpus of Chinese Dynastic Histories: Gender Analysis over Two Millennia. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 785–793. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.