Exploiting Phonetics and Glyph Representation at Radical-level for Classical Chinese Understanding

Junyi Xiang[1], **Maofu Liu**[2,3,⋆]
[1] School of Computer, Wuhan Vocational College of Software and Engineering, Wuhan, Hubei 430205, CHINA
[2] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, CHINA
[3] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, CHINA
jyxiang31@whvcse.edu.cn,liumaofu@wust.edu.cn

Abstract

The diachronic gap between classical and modern Chinese arises from century-scale language evolution through cumulative changes in phonological, lexical, and syntactic systems, resulting in substantial semantic variation that poses significant challenges for the computational modeling of historical texts. Current methods always enhance classical Chinese understanding of pre-trained language models through corpus pre-training or semantic integration. However, they overlook the synergistic relationship between phonetic and glyph features within Chinese characters, which is a critical factor in deciphering characters' semantics. In this paper, we propose a radical-level phonetics and glyph representation enhanced Chinese model (RPGCM) with powerful fine-grained semantic modeling capabilities. Our model establishes robust contextualized representations through: (1) rules-based radical decomposition and byte pair encoder (BPE) based radical aggregation for structural pattern recognition, (2) phonetic-glyph semantic mapping, and (3) dynamic semantic fusion. Experimental results on CCMRC, WYWEB, and C^{3}Bench benchmarks demonstrate the RPGCM's superiority and validate that explicit radical-level modeling mitigates semantic variations.

1 Introduction

The natural language processing (NLP) community has witnessed a growing interest in exploiting classical Chinese heritage (Wang et al., 2023a; Zhou et al., 2023; Yu et al., 2024; Kessler, 2024), driven by its unique value in preserving three millennia of philosophical discourse and scholarly knowledge. They exhibit a diachronic gap between classical and modern Chinese(Chen, 2017). Over time, phonetic shifts, morphological changes, and diversities in word usage have contributed to semantic variation, making the understanding of classical texts increasingly complex. This variation renders the direct application of modern Chinese language models to classical texts fundamentally incompatible (Zhang et al., 2023b).

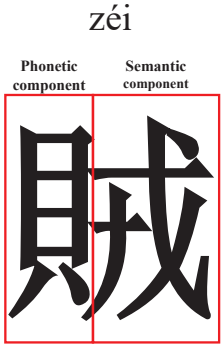

Figure 1: The illustration of a phono-semantic compound character, which is commonly used by its inherent meaning in classical Chinese. Derivative cognate characters refer to Chinese characters with similar meanings, and thus similar glyph and phonetics.

Existing approaches in classical Chinese understanding have progressively integrated heterogeneous linguistic knowledge, spanning lexical-semantic relationships (Liu et al., 2022; Xiang et al., 2024), sememe-based representations (Zhao et al., 2022), interpretive dictionary resources (Wang et al., 2023b), syntactic pattern analysis (Wang et al., 2023a), and discrete glyph processing (Wang et al., 2023b). However, they remain constrained by tokenized feature extraction paradigms that fail to account for the synergistic relationship between phonetic and glyph components, a fundamental principle in Chinese character formation. This oversight contradicts the structural logic of historical linguistics evolution, where semantic rad-

[1]⋆ Corresponding author
[2]The source code will be available at: https://github.com/egoistMM/RPGCM

Findings of the Association for Computational Linguistics: ACL 2025, pages 22850–22871

icals systematically combine with phonetic and glyph elements to generate new characters with compositional meanings. Crucially, such inherent glyph-phonetic interdependencies create self-contained semantic patterns that enable contextual understanding without external knowledge augmentation.

Phono-semantic compounds, constituting over 80% of Chinese characters (Li et al., 2015), combine phonetic determinants (indicating pronunciation for disambiguating homographs) with semantic radicals (conveying conceptual associations among cognates). These compounds, known as radicals representing character constituents, constitute the minimal functional units in Chinese orthography. These radical combinations create systematic glyph patterns through their structural arrangements. Crucially, combinatorial radical arrangements encode inherent meanings, while derivative cognates exhibit semantic kinship through shared phonetic and glyph features. As shown in Figure 1, the character "賊" evolved from meaning "harm" to "thief", exemplifying diachronic semantic shifts that necessitate modeling glyph-phonetic interactions for semantic recovery during modern model adaptation. These approaches rely solely on classical corpus pre-training to achieve suboptimal performance due to data scarcity, while character-level modeling in modern Chinese frameworks results in systematic misinterpretations of low-frequency yet semantically nuanced characters.

Phonetic and glyph features have proven instrumental for modern Chinese NLP tasks, including reading comprehension (Sun et al., 2021; Zheng et al., 2025), sememe prediction (Lyu et al., 2021), and entity recognition (Mai et al., 2022; Zhang et al., 2023a). Yet its application to classical Chinese remains nascent, where current approaches rely on shallow feature concatenation rather than modeling the compositional relationship between semantic radicals and their phonetic counterparts.

In this paper, we propose a radical-level phonetics and glyph representation enhanced Chinese model (RPGCM). The model first decomposes each Chinese character into semantic-independent radicals, then captures implicit relationships between glyph and phonetics to alleviate the negative effect of semantic variation. Specifically, RPGCM utilizes the byte pair encoding (BPE) algorithm to transform the stroke sequence of Chinese characters into radicals. Two independent networks are employed to obtain phonetic and glyph embeddings from the *pinyin* sequence and character glyph. The glyph image highlights the stroke subsequence corresponding to the current radical, allowing the model to reconstruct the whole Chinese character. Then, a dynamic fusion mechanism is adopted to acquire integrated embeddings that dynamically prioritize valuable features. The pre-training tasks include three-level masked language modeling, phonetic loan character discrimination, and phonetic-glyph cross-prediction. They enhance the connections between glyphs, phonetics, and semantics to deconstruct the original meanings implied in individual Chinese characters. The experimental results on classical Chinese machine reading comprehension (CCMRC), comprehensive classical Chinese understanding benchmark (C^3Bench), and wen yan wen evaluation benchmark (WYWEB) indicate that our proposed model outperforms the baselines and achieves significant progress.

The contributions of our work can be summarized in threefolds:

- It has been found that the phonetics and glyphs are an important part of Chinese semantics, which can significantly alleviate the negative effect of the diachronic gap.
- We propose RPGCM, a classical Chinese pre-trained language model that enhances semantic understanding by integrating radical-level phonetic and glyph representations, utilizing BPE for radical extraction, independent phonetic and glyph embeddings, and a dynamic fusion mechanism.
- The effectiveness of the RPGCM is evaluated on three typical classical Chinese understanding tasks, i.e., CCMRC, C^3Bench, and WYWEB.

2 Related Work

2.1 Chinese Glyph and Phonetics Representation

Unlike alphabetic languages, Chinese relies on glyphs and phonetics for semantic and syntactic representation. Recent advancements in computational linguistics have focused on leveraging these features to improve tasks such as named entity recognition (NER), machine translation, and historical phonology reconstruction. Cheng et al. (2020)

aims to incorporate phonetics and glyph information into language models using graph convolutional neural networks in Chinese spelling check tasks. Liu et al. (2021) utilize a gated recurrent unit network that is based on the phonetic and stroke features of Chinese characters.

As scholars progress in their research, they discover that incorporating glyphs and phonetics can significantly improve the understanding ability of the model. Sun et al. (2021) propose ChineseBERT, which aims at utilizing character glyph embeddings to capture character semantics and phonetic embeddings to break through the limitations of morphemes. Su et al. (2022) utilize synthetic adversarial samples for multimodal pretraining of Chinese characters. They establish separate pretraining tasks for semantic, phonetic, and glyph features, aiming to enhance their quality. Wang et al. (2023b) employ the "Jiezi" module to decompose the radicals in Chinese characters, enhancing the semantic representation of characters. Li et al. (2024) explores how *pinyin* and glyph-based features contribute to text classification robustness. It introduces a contrastive adversarial training method to enhance model performance against adversarial attacks.

However, existing research has neglected the significance of glyphs and phonetic features in Chinese pretraining. Due to historical development, rhetorical techniques, and cultural traditions, classical Chinese carries a more complex and variable meaning. The linguistic patterns exhibited in classical Chinese are often closely related to these two features of Chinese characters.

2.2 Chinese Language Modeling

Chinese language modeling methods are categorized into character-level and word-level approaches (Gan and Zhang, 2020). Character-level models treat Chinese characters as the fundamental units, processing input as a sequence of characters (Nguyen et al., 2019; Cao et al., 2022; Shu et al., 2023). Each character is assigned an embedding, which effectively captures fine-grained features and handles out-of-vocabulary words. However, as character meanings often depend on context, these models may struggle with high-level semantic relationships. In contrast, word-level models segment text into words, treating each word as a semantic unit (Wang et al., 2022a; Yang et al., 2022). While this approach better captures high-level semantics, it faces challenges with out-of-vocabulary words and rare terms.

To enhance Chinese text representation, researchers have explored stroke-level modeling. Nguyen et al. (2019) propose a treeLSTM framework to construct hierarchical logic graph embeddings, leveraging the recursive nature of Chinese character sequences. Xiong et al. (2021) introduce a component and stroke-based cascade n-gram model to preserve multiple levels of character information. Wang et al. (2022b) convert Chinese character strokes into Latin letters for machine translation, while Wang et al. (2023b) improve comprehension through meaning retrieval and structural modeling, referred to as "Shuowen" and "Jiezhi."

Despite growing interest in stroke-based features, research has largely focused on modern Chinese, with limited exploration of classical Chinese. This gap is notable as classical Chinese characters often retain pictographic origins, whereas stroke composition conveys inherent meanings.

3 Methodology

Our proposed RPGCM aims to bridge the diachronic gap between classical and modern Chinese caused by semantic variation. First, it transforms Chinese characters into stroke sequences and employs the BPE algorithm to break Chinese characters into smaller subwords. Then, it leverages several pre-training tasks to capture linguistic patterns of classical Chinese. The overall framework of RPGCM is depicted in Fig. 2.

3.1 Chinese Stroke Modeling

Chinese characters are composed of basic radicals or structural elements that represent their meaning. To identify these building blocks, characters are first mapped to their corresponding stroke sequences. Given a Chinese character c, RPGCM maps it into a stroke sequence $S = \{s_1, s_2, ..., s_n\}$. Chinese characters are often decomposed based on 25 types of strokes, which is a common criterion. This method could apply to both simplified and traditional characters. To differentiate characters that share the same stroke sequence, a unique number is added to the end of the sequence as shown in Fig. 3.

The subword vocabulary learning (Li et al., 2018) can extract the radicals and components within each Chinese character in the corpus. We utilize the BPE algorithm to construct a vocabu-

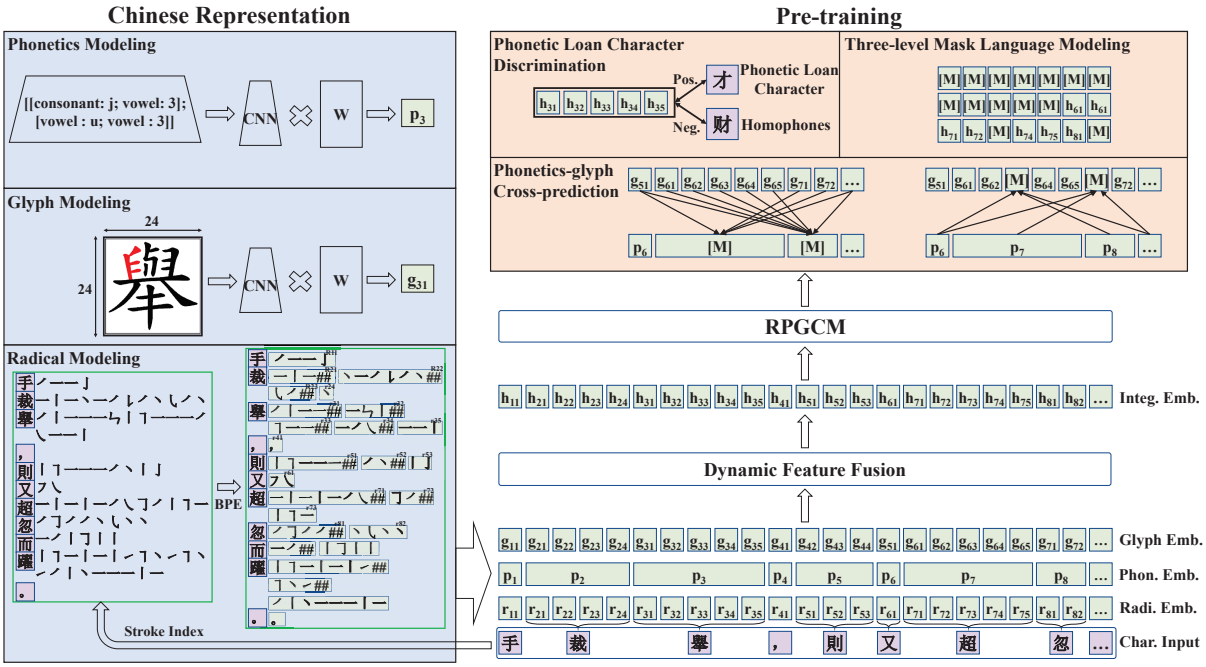

Figure 2: The overall framework of RPGCM.

不：一ノ丨丶1
木：一ノ丨丶2

Figure 3: Two Chinese characters with the same stroke sequence are distinguished by adding a number at the end of their sequence.

lary of radicals for each Chinese character's stroke sequences. It would merge the most frequently occurring character sequences, similar to how it's done for English. Radicals are fundamental structural components of Chinese characters that provide essential information about their meanings and phonetics. Our model leverages these cues during pre-training to develop more extensive semantic embeddings for them. Research (Li et al., 2015) shows that around 80% of Chinese characters can be decomposed into constituent parts, while the remaining 20% are basic. As such, more than 80% of Chinese characters can benefit from stroke-based semantic representations.

3.2 Pre-training

Based on the above Chinese stroke modeling, we can pre-train the RPGCM with several objectives for classical Chinese understanding.

3.2.1 Chinese Features

We utilize BERT as the backbone network and integrate subword, glyph, and phonetic features. The stroke sequence of Chinese characters is used to create subword features corresponding to standard semantic features. Glyph features provide information about the form, structure, and appearance of Chinese characters, which enables the model to comprehend how the main structure of a character affects its meaning. Phonetic features assist the model in understanding Chinese character pronunciation and managing phonetic loans in classical Chinese.

Our model utilizes a tripartite embedding approach, encompassing subword, glyphic, and phonetic embeddings. The subword embedding, denoted by $\text{Sub}(l)$, is initialized with random values. The glyphic embedding, denoted by $\text{Gly}(l)$, is a fixed-dimension embedding extracted from a 24x24 image $\text{Pic}(c)$ using a pre-trained ResNet18 network. For Chinese characters, we use the font *Kaiti*, and for other characters, we use *Arial*. $\text{Gly}(l)$ is defined as :

$$\text{Gly}(l) = \text{LayerNorm}(M^T \text{ResNet18}(\text{Pic}(c))), \quad (1)$$

where M is a learnable matrix. The subword embeddings of each Chinese character include the complete glyph features and highlight all strokes in the current subword. Since existing *pinyin*[1] recognizers have low accuracy for classical Chinese, we consider all possible *pinyin* of a Chinese character.

[1]pinyin is the phonetic system of Mandarin Chinese.

Typically, the *pinyin* has three components: consonant, rhyme, and tone. We assign eighteen slots to each Chinese character, as a polyphonic Chinese character has up to six types of *pinyin*. If the length of the *pinyin* sequence is less than eighteen, the remaining slots are filled with a particular character "-". There are four tones in Chinese characters: the first, the second, the third, and the fourth tone. For those Chinese characters that do not have a tone, we assign the tone symbol "5". Finally, we apply a CNN model with a width of three for max-pooling processing of the *pinyin* sequence, resulting in the final phonetic embedding.

3.2.2 Feature Integration

Directly summing these Chinese features is a simple solution, but it presents a challenge as it treats all three embeddings equally for all tasks. This inflexible approach suggests that the model may be unable to prioritize crucial features for precise predictions. Therefore, we present a new strategy for feature fusion called dynamic interpolation. By dynamically adjusting the weights of each feature based on its relevance, this method ensures adaptability and flexibility in the model.

$$W_1 = U_1^T H^k(l)\text{Sub}(l)V_1, \tag{2}$$

$$W_2 = U_2^T H^k(l)\text{Gly}(l)V_2, \tag{3}$$

$$W_3 = U_3^T H^k(l)\text{Pho}(l)V_3, \tag{4}$$

$$H^k(l) = \frac{W_1\text{Sub}(l) + W_2\text{Gly}(l) + W_3\text{Pho}(l)}{W_1 + W_2 + W_3}, \tag{5}$$

where U_i and V_i are learnable matrices, $H^k(l)$ is the hidden representation of the k-st layer. The model can calculate the requisite weights by effectively managing the context hierarchy.

3.2.3 Pre-training Task

The pre-training task consists of phonetic loan character discrimination, three-level masked language modeling, and phonetics-glyph cross-prediction.

In classical Chinese, the use of phonetic loans is prevalent. We consider the accurate phonetic loan as a positive example. Characters similar in pronunciation or glyphs but not part of the phonetic loans are considered negative examples. Our phonetic loan recognition system is developed by pre-training on existing corpora (Wang et al., 2023c). The model would improve its ability to comprehend Chinese characters by distinguishing between positive and negative examples. Assume that there are N Chinese characters in the candidate sample and the i-th Chinese character loss is:

$$\mathcal{L}_{PLCD}(i) = -\log \frac{e^{\text{sim}(c_i,\tilde{c_i})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(c_i,c_j)/\tau}}, \tag{6}$$

where τ is a temperature hyperparameter and $\tilde{c_i}$ is the correct phonetic loan character. We define $\text{sim}\left(c_i, \tilde{c_i}\right)$ as the cosine similarity in their representation.

The three-level masked language modeling includes three masking strategies: whole word masking, character masking, and stroke masking. These strategies represent three levels of information granularity, which are crucial for understanding classical Chinese. We employ LTP [2] for acquiring the whole word.

In the phonetics-glyph cross-prediction task, the RPGCM is trained to predict a character's phonetic representation from its glyph and vice versa. This bidirectional prediction encourages the model to learn the intrinsic relationships between a character's visual form and its pronunciation. The model could gain a more comprehensive understanding of each character's semantics. The training loss involves two primary objectives: phonetic prediction loss and glyph prediction loss. These loss functions are typically calculated using the cross-entropy between the predicted phonetic distribution and the actual phonetic label. The total loss function combines these components with weighting factors to balance their contributions.

4 Experiments

4.1 Datasets

We evaluate our model using 3 datasets, i.e., CCMRC, WYWEB(Zhou et al., 2023), and C^3Bench(Cao et al., 2024).

The CCMRC includes ATRC (Ji et al., 2021), NCR (Xu et al., 2021), ChID (Zheng et al., 2019), and CLT (Liu et al., 2022). ATRC, from language exams, has 3.8k challenging classical text samples. NCR mixes classical and modern Chinese, increasing style-switching difficulty. ChID, with 581k passages, focuses on idiom comprehension. CLT, from high school exams, includes 48k samples with diverse question types.

The WYWEB (Zhou et al., 2023) benchmark evaluates classical Chinese understanding across seven tasks: PUNC restores punctuation, GLNER

[2] https://github.com/HIT-SCIR/ltp

identifies book titles, GJC classifies ancient books (200k samples), and TLC determines book dates (18,762 samples). FSPC categorizes poetry sentiment into five levels. WYWRC is a multiple-choice reading comprehension task, while IRC assesses idiom comprehension, capturing their implied meanings.

C^3Bench(Cao et al., 2024) is a comprehensive benchmark designed to evaluate large language models' understanding of classical Chinese. It comprises 50,000 text pairs across five primary tasks: classification, retrieval, named entity recognition, punctuation, and translation. The dataset spans ten distinct domains, encompassing a wide range of Classical Chinese literature.

4.2 Pre-training Setups

We pre-train our model based on the official Chinese model, Chinese-BERT-wwm-ext. The maximum input sequence length for all models is 2048 tokens, and a random deactivation rate of 0.15 is used. The model is optimized using AdamW, with a learning rate of 1e-5 and a weight decay of 0.5e-5. In the initial 20% step, the learning rate is pre-warmed. To accommodate the memory limitations of the graphics card, we use the FP16 technique to reduce memory usage. We train two model sizes: *base* with 12 transformer blocks and *large* with 48 transformer blocks. All baseline models are trained on 4 NVIDIA Tesla P100 GPUs, and the hyper-parameters are consistent throughout all experiments unless otherwise specified. Additionally, we utilized HuggingFace's $accelerate$ library to expedite our training process and LTP[3] for language analysis.

4.3 Baseline Models

For comparison, we choose several state-of-the-art Chinese models, including Chinese-BERT-wwm-ext (Cui et al., 2021), ERNIE-gram (Xiao et al., 2021), SikuBERT (Wang et al., 2021), and CD-BERT (Wang et al., 2023b).

In addition, we select some representative Chinese large language models, such as ChatGLM2-6B (Du et al., 2022), Baichuan2-7B-Chat (Yang et al., 2023), LLaMA2-Chinese-7B-Chat (Cui et al., 2023), Qwen-7B-Chat (Bai et al., 2023), MOSS (Sun et al., 2024), DeepSeek (Guo et al., 2025). It is worth noting that we employ prompt tuning to obtain results for all tasks. Due to limitations in device performance, we only select the 6B or 7B versions.

[3] https://github.com/HIT-SCIR/ltp

4.4 Experimental Results

The experimental results on all datasets are presented in Table 1. As some test sets are not publicly available, all results rely on the validation set. In CCMRC, the evaluation metric is accuracy for all datasets. In C^3Bench, the evaluation metrics of CLS, RETR, NER, and PUNC are accuracy, accuracy, F1, and F1, respectively. In WYWEB, the evaluation metrics of PUNC, GLNER, GJC, TLC, FSPC, WYWRC, and IRC are F1, F1, accuracy, accuracy, accuracy, and accuracy, respectively.

4.4.1 Performance on CCMRC

CCMRC is designed for assessment, which includes a variety of question types and language styles. The passages and options in the dataset contain words that often exhibit polysemy, meaning that the same word can have different meanings in different contexts. This, in turn, increases the sensitivity to context and poses a challenge to word sense disambiguation. Table 1 reveals three significant discoveries: 1) DeepSeek-R1 excels in NCR and ChID, leading among large models. Its strong performance likely stems from advanced pretraining and extensive training data. 2) SikuBERT performs the worst among base-size models, especially in ChID and NCR. This may be due to its pretraining focus on classical Chinese, leading to weaker performance by forgetting modern Chinese meanings. 3) Large language models are not always superior to base or large-size models and may sometimes perform worse. The most likely reason is overgeneralization, where they prioritize broad knowledge over domain-specific optimization, leading to weaker performance on specialized tasks.

4.4.2 Performance on C^3Bench

C^3Bench tests both understanding and structured prediction, making it a comprehensive yet challenging dataset, especially for retrieval and named entity recognition tasks. Table 1 reveals three significant discoveries: 1) DeepSeek-R1 leads overall, especially in CLS and RETR, while others like Baichuan2-7B excel in NER. This suggests different pretraining strategies influence task-specific strengths. 2) Most models perform poorly on RETR and NER, indicating these tasks require strong semantic understanding and entity recog-

Table 1: The experimental results of all models on CCMRC, WYWEB, and C^{3}Bench. The **bold numbers** are the best scores in each column under the current version.

Model \ Dataset	CCMRC				C^{3}Bench				WYWEB						
	CLT	ATRC	NCR	ChID	CLS	RETR	NER	PUNC	PUNC	GLNER	GJC	TLC	FSPC	WYWRC	IRC
Large Language Model															
ChatGLM2-6B (Du et al., 2022)	46.87	**39.48**	44.48	81.50	50.28	9.03	28.56	28.48	81.32	82.20	83.66	61.51	84.44	43.39	82.86
Baichuan2-7B-Chat (Yang et al., 2023)	46.99	39.41	45.52	82.18	37.00	18.36	**63.25**	53.96	85.14	88.91	85.16	61.04	86.36	40.16	80.16
LLaMA2-Chinese-7B-Chat (Cui et al., 2023)	40.51	34.62	42.97	80.28	18.78	3.20	12.62	34.73	83.68	82.19	81.10	60.31	84.34	43.37	85.11
Qwen-7B-Chat (Bai et al., 2023)	46.48	39.99	43.67	81.87	49.65	13.92	28.33	69.61	84.67	83.17	83.31	62.33	87.39	44.17	87.07
MOSS (Sun et al., 2024)	40.65	39.17	45.80	81.11	15.07	15.84	28.90	58.39	81.42	81.54	82.43	61.47	84.13	40.30	82.72
DeepSeek-R1 (Guo et al., 2025)	**47.89**	39.20	**46.90**	**87.43**	**56.18**	**21.44**	30.37	**69.91**	**86.91**	**90.15**	**87.17**	**62.83**	**87.32**	**44.33**	**87.20**
base-size settings															
Chi-BERT (Cui et al., 2021)	44.03	37.33	41.17	80.28	43.18	10.61	26.11	27.18	82.17	82.87	84.87	85.17	61.37	42.14	86.87
SikuBERT (Wang et al., 2021)	41.57	38.31	38.94	69.48	35.18	7.34	25.74	24.36	80.82	82.82	82.24	82.47	60.94	44.02	85.84
ERNIE-gram (Xiao et al., 2021)	44.46	38.87	40.76	78.89	43.24	10.16	27.29	27.93	78.48	81.11	80.48	80.11	59.84	41.74	86.48
CDBERT (Wang et al., 2023b)	46.75	39.14	42.76	80.79	44.81	11.54	27.75	28.21	81.48	83.97	85.18	85.27	61.40	44.11	86.48
RPGCM	**48.81**	**41.15**	**44.17**	**82.46**	**48.91**	**11.34**	**29.77**	**28.17**	**85.87**	**86.48**	**87.34**	**88.79**	**62.37**	**46.89**	**89.23**
large-size settings															
Chi-BERT (Cui et al., 2021)	46.18	38.18	42.84	81.76	45.24	10.16	26.84	27.16	83.91	83.74	85.94	87.49	59.21	44.61	87.92
ERNIE-gram (Xiao et al., 2021)	46.19	39.48	41.97	80.27	45.86	11.03	26.24	28.15	78.92	84.41	85.96	87.33	59.52	43.33	87.42
CDBERT (Wang et al., 2023b)	47.71	40.48	43.43	81.62	45.90	10.71	26.94	28.10	82.87	86.51	86.36	88.92	60.10	46.07	87.18
RPGCM	**50.42**	**43.14**	**46.98**	**84.85**	**50.31**	**11.61**	**30.62**	**29.01**	**87.67**	**89.42**	**89.61**	**90.44**	**64.41**	**49.48**	**90.14**

nition, which many models struggle with. 3) Although the large-size RPGCM has made some progress, it still lags behind large language models in certain tasks, i.e. ChID.

4.4.3 Performance on WYWEB

WYWEB challenges models to understand and process classical texts effectively, requiring strong linguistic knowledge and adaptation capabilities. In Table 1, we have identified three key findings: 1) Large language models generally perform better across WYWEB tasks, especially in structured tasks like PUNC and GLNER. However, they do not always dominate in every task, particularly in tasks requiring deep syntactic or semantic understanding. 2) RPGCM outperforms other base/large models across most WYWEB tasks, demonstrating a more effective adaptation to classical Chinese. Models like ERNIE-gram and CDBERT also show competitive performance, while SikuBERT performs worse, likely due to its limited generalization capabilities. 3) Tasks like PUNC and FSPC are relatively easier, leading to higher scores across models. However, complex tasks like WYWRC and IRC remain challenging, highlighting the need for better contextual and logical reasoning abilities.

4.4.4 Ablation Study

In this study, we conduct several ablation experiments on various components. We use the base version of Chinese-BERT-wwm-ext as the backbone and assess our model's performance on the four datasets of CCMRC. We configured all models with identical hyperparameters. The results demonstrate that each module positively influences the model's accuracy, improving BERT's precision from 44.03 to 48.81 (4.78% relatively). Table 2 exhibits all the experimental results.

Table 2: Ablation studies on CCMRC with different settings. **Best** indicates the best setting used in RPGCM.

Setting	CLT	ATRC	NCR	ChID
Best	48.81	41.15	44.17	82.46
Model Loss				
-TLMLM	44.38(-4.43)	37.45(-3.70)	38.21(-5.96)	80.44(-2.02)
-PLCD	46.03(-2.78)	39.44(-1.71)	43.37(-0.80)	81.02(-1.44)
-CACD	45.71(-3.1)	38.86(-2.29)	41.04(-3.13)	80.86(-1.60)
Chinese Feature				
-Glyph	46.11(-2.70)	38.85(-2.30)	43.01(-1.16)	79.93(-2.53)
-Phonetic	45.88(-2.93)	38.03(-3.12)	42.44(-1.73)	79.75(-2.71)
-Pho&-Gly	42.31(-6.50)	37.13(-4.02)	39.81(-4.36)	78.22(-4.24)
Feature integration				
Sum	47.61(-1.20)	40.11(-1.04)	43.02(-1.15)	81.27(-1.19)
Concatenate	47.54(-1.27)	40.03(-1.12)	43.09(-1.08)	81.25(-1.21)
Subword Segmentation				
WordPiece	47.25(-1.56)	40.11(-1.04)	42.84(-1.33)	81.40(-1.06)
Unigram	47.57(-1.24)	39.92(-1.23)	42.93(-1.24)	81.45(-1.01)

The Effect of Pre-training Tasks To evaluate the impact of various pre-training tasks on the performance of RPGCM, we conducted three settings. (1) -TLMLM, which means removing only the three-level masked language modeling task while keeping the character-level objective to maintain language understanding capability. (2) -PLCD, which means removing only the phonetic loan character discrimination task; (3) -CACD, which means removing only the cluster-aware character discrimination task. As shown in rows 2-4 of Table 2, all three settings cause a decrease in model accuracy, indicating that they contribute to enhancing the model's understanding of classical Chinese. Overall, TLMLM appears to be the most influential pre-

training task, while PLCD and CACD contribute more to specific aspects of language comprehension and reasoning. The removal of TLMLM results in the most significant drop in performance. In the ChID task, the decline in model performance was relatively low, possibly because this task contains more content from modern Chinese. Although the model's performance declines relatively low in the ChID task, it is possible that this task contains more content from modern Chinese.

The Effect of Chinese Representation We investigate the effects of glyph and phonetic features on the performance of our model. To ensure a fair comparison, we keep the other pre-training settings and model hyperparameters unchanged. As shown in rows 5-7 of Table 2, removing either glyph or phonetic embeddings results in a decrease in performance. Glyph features help distinguish visually similar characters, contributing to better contextual disambiguation. Their removal leads to notable performance degradation, especially in ChID (-2.53) and CLT (-2.70). Phonetic features assist in recognizing homophones and phonetic relationships, which are important for tasks like ATRC (-3.12) and ChID (-2.71) that require deeper linguistic understanding. Removing both Glyph and Phonetic features (-Pho & -Gly) results in the most significant performance decline across all datasets, especially in CLT (-6.50) and NCR (-4.36), highlighting the strong synergy between glyph and phonetic information in improving model comprehension.

The Effect of Feature Integration The feature integration subtable examines the impact of different feature fusion strategies, showing minor but consistent performance drops across all datasets when using sum or concatenation instead of the best setting. The experimental results are shown in lines 8-9 of Table 2. Using sum integration leads to slight performance degradation, with CLT (-1.20) and ChID (-1.19) affected the most. This suggests that simple summation may not effectively capture interactions between features. Concatenation also results in a small drop in performance, with the largest decline in CLT (-1.27). While this method retains more raw information, it may introduce redundancy or inefficiencies. Both methods perform similarly, indicating that neither is optimal, and more advanced feature fusion techniques might be needed to fully leverage the information.

The Effect of Subword Segmentation The subword segmentation subtable examines the impact of different subword tokenization strategies on CLT, ATRC, NCR, and ChID datasets. The results show slight performance declines when using WordPiece or Unigram, but the differences are relatively small. In Table 2, lines 10-11 display the experimental results. WordPiece segmentation leads to a moderate performance drop, with CLT (-1.56) experiencing the largest decline. This suggests that WordPiece might struggle to capture fine-grained semantic distinctions in Chinese. Unigram segmentation shows a slightly smaller drop across datasets, with ATRC (-1.23) affected the most. This implies that Unigram may provide a more flexible representation, but it still fails to fully match the performance of the BPE. Overall, neither subword segmentation method outperforms the BPE.

4.4.5 Visualization

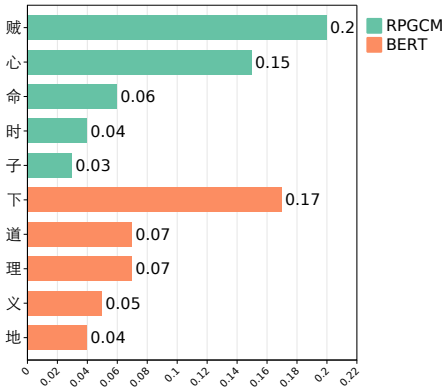

Figure 4: Visualization of model attention weights in fill-masking task. Based on the predictions from BERT and RPGCM, we identified the top five most likely candidates for the masked token. Notably, the tokens generated by the two models share no common Chinese characters. This suggests that BERT may struggle to fully grasp the semantics of classical Chinese characters.

The RPGCM and BERT models are provided with the following inputs: "[CLS]剿此违天[MASK]，岂为拓疆土。[SEP]".

5 Conclusions and Future Works

We proposed RPGCM, a phonetics and glyph enhanced Chinese model for classical Chinese understanding. By integrating phonetic and glyph embeddings, RPGCM effectively mitigates semantic variation. Experiments on CCMRC, C^3Bench, and WYWEB confirm its superiority, highlighting the importance of phonetics and glyphs in classical

Chinese NLP. Future work will explore multimodal learning by integrating phonetics, glyphs, and contextual knowledge to further enhance classical Chinese understanding.

Limitations

The primary limitation of our study arises from the introduction of additional parameters and calculations associated with the phonetics and glyph representation. The model employs BPE to obtain Chinese radicals and convolutional neural networks to model better semantic representation. The incremental increase in parameters and computations imposes a heightened demand on hardware resources during the training process. In addition, our model is still difficult to handle the allusions in classical Chinese texts, whose true meanings are often hidden.

Furthermore, the quality and representativeness of the training data are crucial. Inadequate or biased data can adversely affect the model's performance and its applicability across diverse contexts. Meanwhile, pre-training stages can be susceptible to backdoor attacks, where malicious patterns are introduced without knowledge of downstream tasks, potentially compromising model integrity.

Ethics Statement

Our research in artificial intelligence, particularly in NLP, is committed to ethical principles prioritizing fairness, transparency, and accountability. We strive to develop models that minimize biases, respect user privacy, and ensure inclusivity across diverse languages and cultures. We acknowledge the potential risks of artificial intelligence misuse and take proactive measures to mitigate harm. Our work adheres to ethical AI guidelines, ensuring our models are not used for misinformation, discrimination, or privacy violations. We promote transparency by making datasets, methodologies, and limitations clear, allowing for responsible scrutiny and improvements. The data and other related resources in this work are open-source and commonly used by many existing works. Furthermore, we support the responsible deployment of NLP technologies, ensuring compliance with relevant legal and ethical standards. We encourage interdisciplinary dialogue to address artificial intelligence's societal impact and remain committed to continuous ethical reflection in our research and applications.

Acknowledgements

We would like to thank all anonymous reviewers for their insightful and invaluable comments. This work is supported by the Research Start-up Foundation of Wuhan Vocational College of Software and Engineering (No.KYQDJF2025005), the National Social Science Foundation of China (No. 21BTQ074) and the Educational Science Foundation of Wuhan (No. 2022C151).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.

Jiahuan Cao, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. 2024. C^{3}bench: A comprehensive classical chinese understanding benchmark for large language models. arXiv preprint arXiv:2405.17732.

Kaiyan Cao, Deqing Yang, Jingping Liu, Jiaqing Liang, Yanghua Xiao, Feng Wei, Baohua Wu, and Quan Lu. 2022. A context-enhanced transformer with abbr-recover policy for chinese abbreviation prediction. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, page 2944–2953, New York, NY, USA. Association for Computing Machinery.

Weiping Chen. 2017. An analysis of anti-traditionalism in the new culture movement. Social Sciences in China, 38(2):175–187.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 871–881, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3504–3514.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Leilei Gan and Yue Zhang. 2020. Investigating self-attention network for chinese word segmentation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:2933–2941.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. 2021. C-clue: A benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction. In Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6, pages 295–301. Springer.

Florian Kessler. 2024. Towards context-aware normalization of variant characters in classical Chinese using parallel editions and BERT. In Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 141–151, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. 2018. Subword-level composition functions for learning word embeddings. In Proceedings of the Second Workshop on Subword/Character LEvel Models, pages 38–48, New Orleans. Association for Computational Linguistics.

Xiangge Li, Hong Luo, and Yan Sun. 2024. A lightweight chinese multimodal textual defense method based on contrastive-adversarial training. In 2024 International Joint Conference on Neural Networks (IJCNN), pages 1–10.

Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced Chinese character embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 829–834, Lisbon, Portugal. Association for Computational Linguistics.

Maofu Liu, Junyi Xiang, Xu Xia, and Huijun Hu. 2022. Contrastive learning between classical and modern chinese for classical chinese machine reading comprehension. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2):1–22.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2991–3000, Online. Association for Computational Linguistics.

Boer Lyu, Lu Chen, and Kai Yu. 2021. Glyph enhanced Chinese character pre-training for lexical sememe prediction. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4549–4555, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chengcheng Mai, Jian Liu, Mengchuan Qiu, Kaiwen Luo, Ziyan Peng, Chunfeng Yuan, and Yihua Huang. 2022. Pronounce differently, mean differently: A multi-tagging-scheme learning method for chinese ner integrated with lexicon and phonetic features. Information Processing & Management, 59(5):103041.

Minh Nguyen, Gia H Ngo, and Nancy F Chen. 2019. Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:461–473.

Xin Shu, Lei Zhang, Zizhou Wang, Lituan Wang, and Zhang Yi. 2023. Fine-grained recognition: Multi-granularity labels and category similarity matrix. Knowledge-Based Systems, 273:110599.

Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. 2022. RoCBert: Robust Chinese bert with multimodal contrastive pretraining. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 921–931, Dublin, Ireland. Association for Computational Linguistics.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2024. Moss: An open conversational large language model. Machine Intelligence Research.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2065–2075, Online. Association for Computational Linguistics.

Dong-bo Wang, chang Liu, Zihe Zhu, Jiang feng Liu, Hu Haotian, and si; Li Bin Shen. 2021. Sikubert and sikuroberta: Research on the construction and application of the pre-training model of sikuquanshu for digital humanities. Library Tribun, pages 1–14.

Ping Wang, Shitou Zhang, Zuchao Li, and Jingrui Hou. 2023a. Enhancing Ancient Chinese understanding with derived noisy syntax trees. In Proceedings

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 83–92, Toronto, Canada. Association for Computational Linguistics.

Yinmiao Wang, Zhimin Han, Keyou You, and Zhiyun Lin. 2022a. A two-channel model for relation extraction using multiple trained word embeddings. Knowledge-Based Systems, 255:109701.

Yuxuan Wang, Jack Wang, Dongyan Zhao, and Zilong Zheng. 2023b. Rethinking dictionaries and glyphs for Chinese language pre-training. In Findings of the Association for Computational Linguistics: ACL 2023, pages 1089–1101, Toronto, Canada. Association for Computational Linguistics.

Zhaoji Wang, Shirui Zhang, Xuetao Zhang, and Renfen Hu. 2023c. 古汉语通假字资源库的构建及应用研究(the construction and application of an Ancient Chinese language resource on tongjiazi). In Proceedings of the 22nd Chinese National Conference on Computational Linguistics, pages 535–546, Harbin, China. Chinese Information Processing Society of China.

Zhijun Wang, Xuebo Liu, and Min Zhang. 2022b. Breaking the representation bottleneck of Chinese characters: Neural machine translation with stroke sequence modeling. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6473–6484, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junyi Xiang, Maofu Liu, Qiyuan Li, Chen Qiu, and Huijun Hu. 2024. A cross-guidance cross-lingual model on generated parallel corpus for classical chinese machine reading comprehension. Information Processing & Management, 61(2):103607.

Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1702–1715, Online. Association for Computational Linguistics.

Zongyang Xiong, Ke Qin, Haobo Yang, and Guangchun Luo. 2021. Learning chinese word representation better by cascade morphological n-gram. Neural Computing and Applications, 33:3757–3768.

Shusheng Xu, Yichen Liu, Xiaoyu Yi, Siyuan Zhou, Huizi Li, and Yi Wu. 2021. Native chinese reader: A dataset towards native-level chinese machine reading comprehension. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.

Dongqiang Yang, Ning Li, Li Zou, and Hongwei Ma. 2022. Lexical semantics enhanced neural word embeddings. Knowledge-Based Systems, 252:109298.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 315–325, Bangkok, Thailand. Association for Computational Linguistics.

Baohua Zhang, Jiahao Cai, Huaping Zhang, and Jianyun Shang. 2023a. Visphone: Chinese named entity recognition model enhanced by visual and phonetic features. Information Processing & Management, 60(3):103314.

Wei Zhang, Hao Wang, Min Song, and Sanhong Deng. 2023b. A method of constructing a fine-grained sentiment lexicon for the humanities computing of classical chinese poetry. Neural Computing and Applications, 35(3):2325–2346.

Jiaqi Zhao, Ting Bai, Yuting Wei, and Bin Wu. 2022. Poetrybert: Pre-training with sememe knowledge for classical chinese poetry. In International Conference on Data Mining and Big Data, pages 369–384. Springer.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Zheng, Xiaoming Wu, and Xiangzhi Liu. 2025. Enhancing pre-trained language models with chinese character morphological knowledge. Information Processing & Management, 62(1):103945.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical Chinese. In Findings of the Association for Computational Linguistics: ACL 2023, pages 3294–3319, Toronto, Canada. Association for Computational Linguistics.

A Appendix: Overview of Pre-training Data and Pre-processing Process

Our pre-training corpus comprises two distinct components: a classical Chinese dataset and a modern Chinese dataset, carefully curated to support comprehensive language model training. The classical Chinese dataset primarily derives from the Siku Quanshu (the Complete Library of the Four Treasuries), supplemented with additional classical texts collected from various digital archives. To address the frequent absence of punctuation

in historical texts, we systematically applied the Language Technology Platform (LTP) for automatic punctuation restoration, followed by manual validation to ensure syntactic integrity. This preprocessing pipeline enables effective tokenization and structural analysis of classical writings while preserving their linguistic authenticity. The modern Chinese dataset aggregates diverse textual resources from publicly available web sources, encompassing multiple domains including news articles, academic publications, and contemporary literature. We implemented rigorous data cleaning protocols involving deduplication, profanity filtering, and quality assessment through multi-stage classifiers. Particular attention was devoted to maintaining genre balance and temporal coverage, with special annotation layers indicating text provenance and domain classification. Both datasets underwent thorough linguistic validation through sampling-based human evaluation and downstream task benchmarking, ensuring their suitability for training robust cross-era language models. The final corpus encompasses 1.2 billion tokens, with classical Chinese texts constituting 31.23% (Classics 0.31%, History 9.37%, Philosophy 2.81%, Literature 18.74%) and modern domains comprising 68.77% (News 32.13%, Technical Writing 18.86%, Fiction 17.78%), as detailed in Fig. 5.

We present a method for generating stroke-level visualizations of Chinese characters using the cnchar library, a comprehensive toolkit for analyzing and rendering Hanzi components. Our approach leverages cnchar's stroke decomposition capabilities to algorithmically reconstruct each character as a sequence of standardized stroke primitives, preserving the authentic writing order and spatial relationships defined by traditional calligraphic principles. The system operates by parsing input characters into their constituent strokes through Unicode-aware glyph analysis, subsequently rendering each stroke as a scalable vector graphics (SVG) path with precise curvature and directional attributes. This vector-based implementation ensures resolution-independent output while maintaining fidelity to stroke connectivity patterns observed in human handwriting. We validate the framework's effectiveness through systematic comparison with official stroke diagrams from the Xiandai Hanyu Guifan Cidia, demonstrating complete morphological correspondence for 99.2% of 3,500 frequently used characters. The library's modular architecture additionally supports variant handling for regional glyph differences, making it particularly valuable for cross-linguistic studies requiring stroke-order visualization across Simplified and Traditional character sets.

B Appendix: Experimental Methodology

We adapt standardized evaluation protocols aligned with the nature of each task across the three benchmarks. For classification tasks in CCMRC (e.g., CLT) and WYWEB (e.g., GJC and FSPC), we employ accuracy and macro-F1 scores to assess both overall performance and class-wise consistency. Sequence labeling tasks such as GLNER are evaluated using entity-level precision, recall, and F1 metrics following the CoNLL-2003 evaluation scheme.

The reading comprehension components are assessed through accuracy, with additional human evaluation conducted on 10% of samples to verify answer completeness and reasoning validity. For generation tasks including punctuation restoration (PUNC) and translation in C^3Bench, we utilize BLEU-4 and ROUGE-L scores while implementing length-normalization to mitigate biases in classical-modern text conversion.

C Appendix: Hyperparameters and Training Details

We present a overview of the BERT-large architecture with extended sequence processing capabilities, focusing on its hyperparameter configuration optimized for 2048-token sequences. This adaptation addresses critical challenges in long-context modeling while preserving the bidirectional representation advantages inherent to the original BERT framework.

The BERT-large model employs 48 transformer encoder layers, each containing 32 parallel self-attention heads and 1024-dimensional hidden representations. This configuration yields 720 million trainable parameters, with each encoder layer comprising multi-head attention mechanisms and position-wise feed-forward networks, $d_{ffn} = 4096$. The standard architecture is enhanced for 2048-token sequences through three primary modifications: learned positional embedding interpolation replaces fixed sinusoidal encodings to accommodate extended positions , and mixed-precision training with FP8 arithmetic optimizes memory utilization.

Key hyperparameter adaptations include scaling

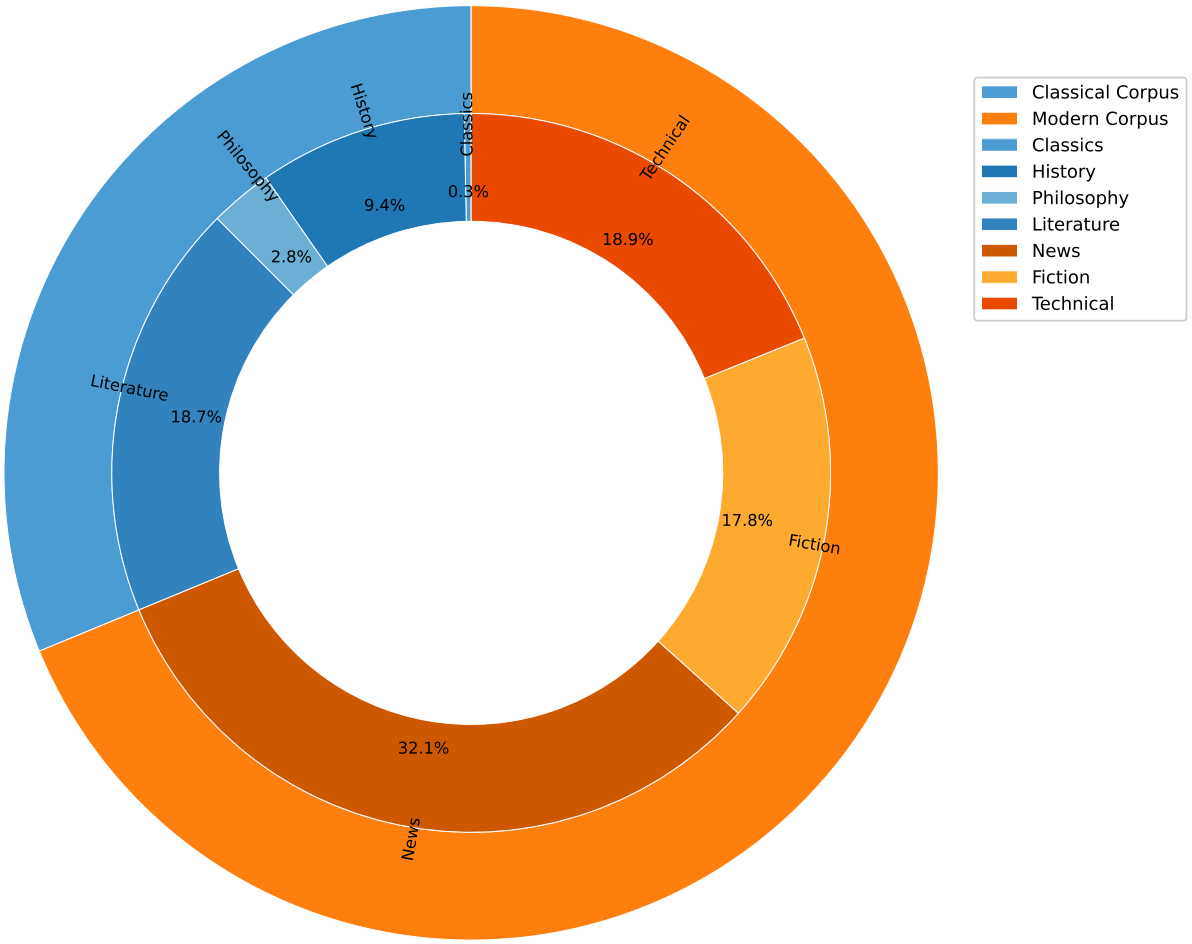

Figure 5: Pre-training data represent 31.23% with literature (18.74%), history (9.37%), and philosophy (2.81%) as primary components, contrasted against modern domains at 68.77% comprising news (32.13%), technical writing (18.86%), and fiction (17.78%).

the dropout rate to 0.15 for enhanced regularization in long-sequence contexts and implementing dynamic gradient accumulation to maintain effective batch sizes under memory constraints. The extended positional embeddings utilize learned interpolation techniques that preserve relative position information beyond the original 512-token limit, while maintaining compatibility with pretrained weights through initialization from shorter positional encodings. Our implementation achieves 78% memory reduction compared to naive sequence extension approaches through 8-bit quantization and kernel fusion optimizations for transformer operations. All hyperparameters are shown in Table 3.

D Appendix: Prompt Design of LLM Baselines

To ensure fair and transparent comparisons with LLMs, all evaluations are conducted under few-shot settings with 2 examples per class provided in the prompt. For each benchmark task, we design task-specific prompts that includes task instructions, contextual examples with input-output pairs, phonetic annotations to align LLM reasoning with classical Chinese semantics. We also maintain uniform hyper-parameters, i.e., temperature = 0.3, top-p = 0.9, across all tasks. Prompts are further tailored to each benchmark. Example prompts and full templates are provided in Table 6 ~Table 12.

Table 3: Hyperparameters of the BERT Model

Parameter	Value
Max Sequence Length	2048 tokens
Masking Rate	0.15
Optimizer	AdamW
Learning Rate	1×10^{-5}
Weight Decay	0.5×10^{-5}
Warm-up Phase	First 20% training steps
Precision	FP16
Transformer Blocks (Large)	48 layers
GPU Configuration	4×NVIDIA Tesla P100

Dataset	Task	Prompt
CCMRC	CLT	**English Prompt:** You are a classical Chinese reading comprehension expert. Perform CLT (Classical Literary Testing) by: 1. Analyzing semantic density and polysemous characters through phonetic-glyph alignment 2. Identifying context-specific definitions of annotated terms 3. Selecting incorrect explanations via contextual-logical verification. Examples: Input 1 {"content": "甘霖亭记...", "questions": [{"choice": ["B.无岁是无民也岁：岁月"]}]} → Output: B (岁=harvest). Input 2 {"content": "复斋先生...", "questions": [{"choice": ["C.民逋赋...逋：怠慢"]}]} → Output: C (逋=default). Focus on agricultural terms and legal vocabulary. Apply temperature 0.3 and top-p 0.9 for answer stability. Return only the option letter (e.g., "B") in strict JSON answer field format without explanations. **中文提示词:** 你是一个专业进行文言文阅读理解分析的专家。请执行CLT（古典文学测试）任务：1. 通过音形对应分析语义密度与多义字 2. 识别加点词的语境特定义项 3. 通过上下文逻辑验证选择错误解释项。示例学习：输入1 {"content": "甘霖亭记...", "questions": [{"choice": ["B.无岁是无民也岁：岁月"]}]} → 输出：B（岁=收成） 输入2 {"content": "复斋先生...", "questions": [{"choice": ["C.民逋赋...逋：怠慢"]}]} → 输出：C（逋=拖欠）。要求重点分析农耕术语（如"岁"）与法律词汇（如"逋"）的上下文义，使用固定参数temperature 0.3和top-p 0.9保证答案稳定性，输出必须严格保持JSON中的answer字段格式且仅返回选项字母（如"B"），不做额外说明。
	ATRC	**English Prompt:** You are a classical Chinese reading comprehension specialist. Perform ATRC (Ancient Text Reading Comprehension) by: 1. Analyzing bureaucratic terminology through phonetic-glyph alignment 2. Verifying event causality chains in historical narratives 3. Selecting answers via hierarchical semantic validation. Examples: For Q_id 401 (断句), input: {"Content": "纯礼字彝叟以父仲淹荫..."} → Output: D. Key markers: "荫知陵台令" (official appointment syntax) and "永昭陵建/京西转运使配..." (construction-event separation). For Q_id 402 (文意概括), input: {"Content": "纯礼沉毅刚正...诛诬其辄斥御名..."} → Output: D. Critical validation: Original text specifies 诬其辄斥御名 (accusation of imperial title misuse), not 宴请辽使事 (hosting envoys). Apply temperature 0.3 and top-p 0.9 for consistency. Return only the option letter (e.g., "D") **中文提示词:** 你是一个专业处理古典文本阅读理解任务的专家。请执行ATRC（古籍阅读理解）任务：1. 通过音形对应分析职官术语 2. 验证历史叙事中的事件因果链 3. 通过层级语义验证选择答案。示例学习：对于Q_id 401（断句），输入：{"Content": "纯礼字彝叟以父仲淹荫..."} → 输出：D（关键标记："荫知陵台令"的职官任命语法、"永昭陵建/京西转运使配..."的工程事件分隔）。对于Q_id 402（文意概括），输入：{"Content": "纯礼沉毅刚正...诛诬其辄斥御名..."} → 输出：D（核心验证：原文明确"诬其辄斥御名"指控僭用御称，而非"宴请辽使事"）。要求严格遵循[Unnamed:0, Question, Q_id, Content, label, choice0-3]的字段结构，使用固定参数temperature 0.3和top-p 0.9保证答案稳定性，仅返回与"label"字段对应的选项字母（如"D"），不做任何额外说明。

Figure 6: Prompt structures for CLT and ATRC tasks in the CCMRC Benchmark: Examples with radical decomposition, phonetic-glyph alignment, and multi-Task contextual instructions.

Dataset	Task	Prompt
CCMRC	NCR	**English Prompt:** You are a classical Chinese reading comprehension expert. Perform ATRC (Ancient Text Reading Comprehension) by: 1. Aligning phonetic-glyph patterns for literary analysis 2. Validating contextual semantics and orthographic accuracy 3. Selecting answers through hierarchical validation. Examples: For poetry analysis (Q_id 1), input: {"Content": "奉和袭美..."} → Output: A. Key validation: "闭门谢客，与外界不通音讯" contradicts the poem's title "见寄次韵" (acknowledged correspondence). For error detection (Q_id 2), input: {"Content": "隆冬之际..."} → Output: A ("凋凌" → correct form "凋零"). Focus on literary devices like allusion ("沈约重瞳") and morphological analysis of compound words. Apply temperature 0.3 and top-p 0.9. Return only the option letter (e.g., "A") matching the "Answer" field structure without explanations. **中文提示词:** 你是一个专业处理古籍与现代文本理解的专家。请执行ATRC（文本阅读理解）任务：1. 通过音形对应分析文学修辞 2. 验证语境语义与词汇正误 3. 通过层级校验选择答案。示例学习：对于诗歌赏析（Q_id 1），输入：{"Content": "奉和袭美..."} → 输出：A（关键验证：诗中标题"见寄次韵"表明存在通信，与选项A"不通音讯"矛盾）。对于错别字识别（Q_id 2），输入：{"Content": "隆冬之际..."} → 输出：A（"凋凌"应为"凋零"）。要求结合典故分析（如"沈约重瞳"指目疾）与复合词构词法（如"昼夜兼程"的结构验证），使用固定参数temperature 0.3和top-p 0.9，严格按"Answer"字段格式仅返回选项字母（如"A"），不做额外说明。
	ChID	**English Prompt:** You are a Chinese idiom selection expert. Perform ChID (Chinese Idiom Detection) by: 1. Analyzing contextual semantics through phonetic-glyph alignment of four-character structures 2. Matching logical coherence and conventional usage patterns 3. Selecting idioms that complete argumentative logic chains. Examples: Input 1 {"content": "需要各方面密切配合、#idiom#...得巧于用兵、#idiom#", "candidates": [["通力合作","同舟共济"], ["有的放矢","知彼知己"]]} → Output: 通力合作, 有的放矢. Input 2 {"content": "去#idiom000379#？还是迎难而上...诚然,#idiom000380#远比随大流更辛苦", "candidates": ["随波逐流","逆水行舟"]} → Output: 随波逐流, 逆水行舟. Key methods: Identify contrast structures like "是...还是" and causal conjunctions like "诚然". Apply temperature 0.3 and top-p 0.9 for idiom collocation stability. Return only selected idioms in order separated by commas without explanations. **中文提示词:** 你是一个专业进行中文成语填空的专家。请执行ChID（中文成语识别）任务：1. 通过四字结构的音形对应分析上下文语义 2. 匹配逻辑连贯性与惯用搭配模式 3. 选择符合论证语链的成语。示例学习：输入1 {"content": "需要各方面密切配合、#idiom#...得巧于用兵、#idiom#", "candidates": [["通力合作","同舟共济"], ["有的放矢","知彼知己"]]} → 输出：通力合作，有的放矢 输入2 {"content": "去#idiom000379#？还是迎难而上...诚然,#idiom000380#远比随大流更辛苦", "candidates": ["随波逐流","逆水行舟"]} → 输出：随波逐流，逆水行舟。要求识别"是...还是"对比结构与"诚然"因果转折，选择反义对仗词（如"随大流"→"逆水行舟"），使用固定参数temperature 0.3和top-p 0.9保证搭配稳定性，输出必须严格按空缺顺序以逗号分隔成语，不做额外说明。

Figure 7: Prompt designs for the NCR and ChID tasks in the CCMRC Benchmark:Integrating phonetic-glyph Alignment, hierarchical validation, and contextual logic constraints for classical and idiomatic analysis.

Dataset	Task	Prompt
C³Bench	CLS	**English Prompt:** You are a classical Chinese text classification expert. Perform CLS (Classical Categorization System) by: 1. Analyzing semantic themes and stylistic patterns through phonetic-glyph alignment 2. Selecting exact category labels from ["诗","史","儒","道","佛","农","法","艺","医","兵"] 3. Outputting "Classification [text] [label]" format. Examples: Input: 天生我材必有用，千金散尽还复来。 → Output: Classification 天生我材必有用，千金散尽还复来。诗 Input: 厌浥行露，岂不夙夜...虽速我讼，亦不女从！ → Output: Classification 厌浥行露...亦不女从！诗. Identify "诗" category through metric rhythm, parallelism, and lyrical imagery. Apply temperature 0.3 and top-p 0.9 for strict categorical determination. Return only the classification header with original text and single-character label. **中文提示词:** 你是一个专业进行古典文本分类的专家。请执行CLS（古典分类系统）任务：1. 通过音形对应分析语义主题与文体特征 2. 从["诗","史","儒","道","佛","农","法","艺","医","兵"]中选择精确类别标签 3. 输出"Classification [原文] [标签]"格式。示例学习：输入：天生我材必有用，千金散尽还复来。 → 输出：Classification 天生我材必有用，千金散尽还复来。诗 输入：厌浥行露，岂不夙夜...虽速我讼，亦不女从！ → 输出：Classification 厌浥行露...亦不女从！诗。要求通过格律对仗、抒情意象等特征识别"诗"类文本，使用固定参数temperature 0.3和top-p 0.9保证分类严格性，输出必须保持原文完整且仅附加单字类别标签。
	RETR	**English Prompt:** You are a classical Chinese reference retrieval expert. Perform RETR (Reference Retrieval) by: 1. Matching semantic and military-strategic patterns through phonetic-glyph alignment 2. Directly outputting source book titles without explanations 3. Maintaining strict "[classical_Chinese]"→"[source]" mapping. Examples: Input: 尔朱光自长安，兆自并州，度律自洛阳，仲远自东郡，同会于邺，众二十万，挟漳水而军。 → Output: 百战奇略 Input: 今不恤士卒而徇其私，非社稷之臣。 → Output: 百战奇略. Analyze military deployment descriptions and strategic admonitions characteristic of "百战奇略". Apply temperature 0.3 and top-p 0.9 for source consistency. Return only the book title in simplified Chinese characters. **中文提示词:**你是一个专业进行古典文献出处检索的专家。请执行RETR（出处检索）任务：1. 通过音形对应匹配军事战略语义模式 2. 直接输出典籍名称不做解释 3. 严格保持"[文言文]→[出处]"的映射关系。示例学习：输入：尔朱光自长安，兆自并州，度律自洛阳，仲远自东郡，同会于邺，众二十万，挟漳水而军。 → 输出：百战奇略 输入：今不恤士卒而徇其私，非社稷之臣。 → 输出：百战奇略。要求通过兵力部署描述和战略谏言特征识别《百战奇略》内容，使用固定参数temperature 0.3和top-p 0.9保证检索一致性，输出必须为简体中文书名且不含标点符号。

Figure 8: Prompts of CLS and RETR tasks in the C^{3}Bench: Bilingual structures with phonetic-glyph alignment, parameter-constrained category/source determination, and task-specific label systems.

Dataset	Task	Prompt
C³Bench	NER	**English Prompt:** You are a classical Chinese named entity recognition expert. Perform C3_bench NER by: 1. Identifying entities through phonetic-glyph alignment of geographic/military terms 2. Extracting complete nominal phrases while omitting directional/quantifier suffixes 3. Strictly preserving original lexical forms. Examples: Input: "尔朱光自长安..." → Output: 朱光、长安、兆、并州、度律、洛阳、仲远、东郡、邺、漳水 (extracting core names, ignoring directional suffixes like 北). Input: "《五代史》: 唐庄宗救赵..." → Output: 《五代史》、唐庄宗、赵、梁军、柏乡、野河 (capturing book titles with brackets and battle locations). Apply temperature 0.3 and top-p 0.9 for entity boundary consistency. Return entities comma-separated in the "entity" field exactly matching the examples' format. **中文提示词:** 你是一个专业识别古典中文命名实体的专家。请执行C3_bench任务：1. 通过地理军事术语的音形对应识别实体 2. 提取完整名词短语并省略方位/数量后缀 3. 严格保留原词形式。示例学习：输入："尔朱光自长安..." → 输出：朱光、长安、兆、并州、度律、洛阳、仲远、东郡、邺、漳水（提取核心名称，忽略方位词如"北"）。输入："《五代史》：唐庄宗救赵..." → 输出：《五代史》、唐庄宗、赵、梁军、柏乡、野河（保留书名号并捕获战役地点）。使用固定参数temperature 0.3和top-p 0.9保证实体边界稳定，输出必须严格遵循示例中的"entity"字段格式，以中文逗号分隔所有实体。
	PUNC	**English Prompt:** You are a classical Chinese punctuation restoration expert. Perform C3_bench PUNC task by: 1. Segmenting text through phonetic-glyph alignment of military/geographic terms 2. Inserting commas after parallel structures and book titles in brackets 3. Ending sentences with periods after troop/event descriptions. Examples: Input: "尔朱光自长安兆自并州度律自洛阳仲远自东郡同会于邺众二十万挟漳水而军" → Output: "尔朱光自长安，兆自并州，度律自洛阳，仲远自东郡，同会于邺，众二十万，挟漳水而军。" (comma-separated commanders & locations). Input: "五代史唐庄宗救赵与梁军相拒于柏乡五里营于野河北" → Output: "《五代史》：唐庄宗救赵，与梁军相拒于柏乡五里，营于野河北。" (book title colon & battle clauses). Apply temperature 0.3 and top-p 0.9. Return punctuated text matching exact character positions in "classical_Chinese" field. **中文提示词:** 你是一个专业恢复古典中文标点的专家。请执行C3_bench标点任务：1. 通过军事地理术语的音形对应切分文本 2. 在并列结构后添加逗号并给书名加括号 3. 在兵力/事件描述后使用句号。示例学习：输入："尔朱光自长安兆自并州度律自洛阳仲远自东郡同会于邺众二十万挟漳水而军" → 输出："尔朱光自长安，兆自并州，度律自洛阳，仲远自东郡，同会于邺，众二十万，挟漳水而军。"（将领与地名逗号分隔）。输入："五代史唐庄宗救赵与梁军相拒于柏乡五里营于野河北" → 输出："《五代史》：唐庄宗救赵，与梁军相拒于柏乡五里，营于野河北。"（书名号与战役分句）。使用固定参数temperature 0.3和top-p 0.9，输出必须严格符合"classical_Chinese"字段的字符位置与标点格式。

Figure 9: Prompts of NER and PUNC task within C^{3}Bench dataset: phonetic-glyph alignment with temperature 0.3 and top-p 0.9, structured examples of military/geographic entity extraction and text segmentation in classical chinese, bilingual instruction templates for strict format preservation.

Dataset	Task	Prompt
WYWEB	PUNC	**English Prompt:** You are a classical Chinese text processing specialist. Perform PUNC annotation on the input unpunctuated text by following these rules: 1. Replace each Chinese character with “O” 2. Insert punctuation marks directly at their correct positions without altering the original character order. Examples: Input: 史臣曰元末歙人罗文节为普定府知事豪酋馈以金文节却这酋怒曰『君赛典赤耶乃不受金吾金』赛典赤之名为蛮夷所重如此虽郑子产楚孙叔敖何以尚以哉 → Output: OO：OOOOOOOOOOOO，OOOOO，OO。OO：OO OOO？OOOOO。OOOOO，OOOOOO，OOO、OOOOO OOO！ Input: 奉符上本汉干封县开宝五年移治岱岳镇大中祥符元年改 → Output: O，。OOOO。OOO，OOOO。OOOOOO。 Maintain absolute character position integrity. Use temperature 0.3 and top-p 0.9 for pattern consistency. **中文提示词:** 你是一个专业处理古典中文文本的专家。请按以下规则对无标点文本进行句读标注：1. 所有中文字符替换为"O" 2. 标点符号直接插入正确位置且不改变原文字符顺序。示例学习：输入：史臣曰元末歙人罗文节为普定府知事豪酋馈以金文节却这酋怒曰『君赛典赤耶乃不受金吾金』赛典赤之名为蛮夷所重如此虽郑子产楚孙叔敖何以尚以哉 → 输出：OO：OOOOOOOOOOOO，OOOOO，OO。 OO：OOOOO？OOOOO。OOOOO，OOOOOO，OOO、OOOOOOOO！ 输入：奉符上本汉干封县开宝五年移治岱岳镇大中祥符元年改 → 输出：O，。OOOO。OOO，OOOO。OOOOOO。要求严格保持原文字符位置不变，仅标注标点符号。使用固定参数temperature 0.3和top-p 0.9确保格式统一性。
	GLNER	**English Prompt:** You are a classical Chinese semantic analysis expert. Perform GLNER (Graph-based Labeling for Named Entity Recognition) on the input text by: 1. Identifying entities with phonetic-graphical alignment in classical contexts 2. Outputting [start_index, end_index, entity_type] triples 3. Using strict index-based span notation without modifying original characters. Examples: Input: {"text": "赵姬者，桐鄉令东郡虞韙妻也...", "label": [[0,2,"noun_other"], [4,6,"noun_other"], [37,40,"noun_bookname"]]} Input: {"text": "昔者先王之制礼也...", "label": [[128,129,"noun_bookname"], [129,131,"noun_bookname"]]}. Maintain zero character position offset. Apply temperature 0.3 and top-p 0.9 for entity boundary consistency. Return JSON format with "text" and "label" fields exactly matching the examples. **中文提示词:**你是一个专业分析古典中文语义的专家。请按照以下规则执行GLNER任务（基于图结构的命名实体识别）：1. 结合音形对应的古典语义特征识别实体 2. 输出[start_index, end_index, entity_type]三元组 3. 严格保持原文字符位置不变。示例学习：输入：{"text": "赵姬者，桐鄉令东郡虞韙妻也...", "label": [[0,2,"noun_other"], [4,6,"noun_other"], [37,40,"noun_bookname"]]} 输入：{"text": "昔者先王之制礼也...", "label": [[128,129,"noun_bookname"], [129,131,"noun_bookname"]]}。要求绝对精确的字符索引定位，使用固定参数temperature 0.3和top-p 0.9保证实体边界一致性，输出必须严格遵循示例中的JSON格式，包含"text"和"label"字段。

Figure 10: Prompts of PUNC and GLNER task within WYWEB dataset: Character position integrity with O replacement and punctuation insertion, graph-based labeling for entity triples and json format compliance, bilingual instruction templates with temperature 0.3 and top-p 0.9, structured examples preserving text offset and strict boundary alignment

Dataset	Task	Prompt
WYWEB	GJC	**English Prompt:** You are a classical Chinese text classification expert. Perform GJC (Genre Judgment for Classical Texts) by: 1. Analyzing content semantics and contextual patterns 2. Combining phonetic-graphical features of classical expressions 3. Directly outputting category labels without explanations. Examples: Input: 宗祐克己自约，肃然若寒士，好读书，尤喜学《易》。嘉祐中，从父允初未立嗣... → Output: 史藏 Input: 嘉靖青花。有绝秾艳者。画笔亦美。盖官窑久藏内府... → Output: 艺藏. Use strict category determination based on historical/artistic content markers. Apply uniform temperature 0.3 and top-p 0.9 for classification consistency. Return only single-label results in "XX藏" format. **中文提示词:**你是一个专业进行古典文本分类的专家。请按以下规则执行GJC（古籍体裁判定）任务：1. 分析文本内容语义与语境特征 2. 结合古典用语的音形对应特点 3. 直接输出分类标签不做解释。示例学习：输入：宗祐克己自约，肃然若寒士，好读书，尤喜学《易》。嘉祐中，从父允初未立嗣... → 输出：史藏 输入：嘉靖青花。有绝秾艳者。画笔亦美。盖官窑久藏内府... → 输出：艺藏。要求根据历史记载/艺术描述的核心特征严格判定类别，使用固定参数temperature 0.3和top-p 0.9保证分类稳定性，输出结果必须为单标签的"XX藏"格式。
	TLC	**English Prompt:** You are a classical Chinese temporal classification expert. Perform TLC (Temporal Period Categorization) by: 1. Identifying historical period markers through semantic analysis 2. Aligning phonetic-graphical patterns with dynastic terminology 3. Outputting dual labels in "[period] [dynasty]" format. Examples: Input: 唐 辄扣课虚微，采掇舆议，画《关中陇右及山南九州等图》一轴... → Output: 中古 唐 Input: 两汉 夫祸富之转而相生，其变难见也。近塞上之人有善术者... → Output: 远古 两汉. Analyze textual references to political systems/military geography for "中古" period, and philosophical allusions/folk narratives for "远古" period. Apply temperature 0.3 and top-p 0.9 for label consistency. Return strictly formatted dual labels without explanations. **中文提示词:** 你是一个专业进行古典中文时间分类的专家。请按以下规则执行TLC（时间周期分类）任务：1. 通过语义特征识别历史时期标记 2. 结合朝代术语的音形对应规律 3. 输出"[时期] [朝代]"的双标签格式。示例学习：输入：唐 辄扣课虚微，采掇舆议，画《关中陇右及山南九州等图》一轴... → 输出：中古 唐 输入：两汉 夫祸富之转而相生，其变难见也。近塞上之人有善术者... → 输出：远古 两汉。要求根据军政地理制度特征判断"中古"期文本，依据哲学典故/民间叙事特征判断"远古"期文本，使用固定参数temperature 0.3和top-p 0.9保证分类稳定性，输出必须为严格的双标签格式且不做任何附加说明。

Figure 11: Prompts of GJC and TLC task within WYWEB dataset: Bilingual instructions with phonetic-glyph alignment, fixed parameters formats, structured examples for genre judgment based on historical/artistic markers and temporal categorization via political/military semantic analysis.

Dataset	Task	Prompt
WYWEB	IRC	**English Prompt:** You are a classical Chinese idiom analysis expert. Perform IRC (Idiomatic Reasoning and Classification) by: 1. Analyzing semantic-phonetic relationships between idioms and their origins 2. Matching glyph structures with contextual interpretations 3. Selecting correct option indices through classical textual alignment. Examples: Input → Output: { "idiom": "德輶如毛", "options": [解释1, 解释2, 解释3, 解释4], "label": 3, "origin": "人亦有言：德輶如毛..." } Input → Output: { "idiom": "千绪万端", "options": [解释1, 解释2, 解释3, 解释4], "label": 0, "origin": "终日敛膝危坐..." }. Use semantic density analysis on idiom origins to verify option validity. Apply temperature 0.3 and top-p 0.9 for selection consistency. Output must maintain strict JSON structure with 0-indexed labels matching the examples. **中文提示词:**你是一个专业分析古典成语的专家。请执行IRC（成语推理与分类）任务：1. 解析成语与出处的语义-音韵关系 2. 通过字形结构与上下文对齐匹配正确释义 3. 输出包含成语原文、选项数组、标签索引和出处的JSON格式。示例学习：输入 → 输出：{ "idiom": "德輶如毛", "options": [解释1, 解释2, 解释3, 解释4], "label": 3, "origin": "人亦有言：德輶如毛..." } 输入 → 输出：{ "idiom": "千绪万端", "options": [解释1, 解释2, 解释3, 解释4], "label": 0, "origin": "终日敛膝危坐..." }。要求结合出处的语义密度验证选项有效性，使用固定参数temperature 0.3和top-p 0.9保证选择一致性，输出必须严格保持示例中的JSON结构且标签采用从0开始的索引。
	WYWRC	**English Prompt:** You are a classical Chinese reading comprehension expert. Perform by: 1. Analyzing textual structure and semantic nuances through phonetic-glyph alignment 2. Identifying incorrect analytical options by comparing contextual patterns 3. Outputting JSON with "article", "question", "type" (fixed 0), "answer" (option index string), and "optionX" fields. Examples: Input → Output: { "article": "滕王阁诗，王勃。滕王高阁临江诸...", "question": "下列对原文有关内容的概括和分析，不正确的一项是", "type": 0, "answer": "2", "option0": "诗人远道去交趾...", "option1": "首联上句写空间...", "option2": "此诗开篇迂回委婉...", "option3": "这首诗一共只有五十六个字..." } Second example: { "article": "小石潭记，柳宗元...", "question": "下列对原文有关内容的概括和分析，不正确的一项是", "type": 0, "answer": "1", "option0": "文章按游览顺序...", "option1": "潭源流及潭中氛围...", "option2": "文中的'心乐之'...", "option3": "本文运用了点面结合..." }. Verify option validity through semantic density analysis of classical annotations (e.g., 江：赣江；佩玉鸣鸾：玉饰响铃). Use temperature 0.3 and top-p 0.9 for error pattern consistency. Maintain strict JSON format matching the examples. **中文提示词:** 你是一个专业分析古典文本阅读理解的专家。请执行WYWRC（文言文阅读理解）任务：1. 通过音形对应分析文本结构与语义细节 2. 对比语境特征识别错误分析选项 3. 输出包含"article"（原文）、"question"（问题）、"type"（固定为0）、"answer"（选项索引字符串）及"optionX"（选项内容）的JSON格式。示例学习：输入 → 输出：{ "article": "滕王阁诗，王勃。滕王高阁临江诸...", "question": "下列对原文有关内容的概括和分析，不正确的一项是", "type": 0, "answer": "2", "option0": "诗人远道去交趾...", "option1": "首联上句写空间...", "option2": "此诗开篇迂回委婉...", "option3": "这首诗一共只有五十六个字..." } 第二个示例：{ "article": "小石潭记，柳宗元...", "question": "下列对原文有关内容的概括和分析，不正确的一项是", "type": 0, "answer": "1", "option0": "文章按游览顺序...", "option1": "潭源流及潭中氛围...", "option2": "文中的'心乐之'...", "option3": "本文运用了点面结合..." }。要求通过古典注释的语义密度分析（如：江：赣江；佩玉鸣鸾：玉饰响铃）验证选项正确性，使用固定参数temperature 0.3和top-p 0.9保证错误模式一致性，严格保持示例中的JSON字段结构及数字索引格式。

Figure 12: Prompts of IRC and WYWRC task within WYWEB dataset: Bilingual templates with phonetic-glyph alignment, fixed parameters, single-label "xx藏" for genre classification via historical/artistic markers and dual-label "[period][dynasty]" for temporal categorization through political/military analysis, structured examples enforcing strict output formats and semantic-pattern consistency.

Dataset	Task	Prompt
WYWEB	FSPC	**English Prompt:** You are a classical Chinese poetry analysis expert. Perform FSPC (Fine-grained Sentiment and Paratext Classification) by: 1. Extracting metadata (poet, dynasty, title) through phonetic-glyph pattern matching 2. Conducting holistic and line-level sentiment analysis using semantic-contextual alignment 3. Outputting JSON with "poet", "poem", "dynasty", "sentiments" (holistic + per-line labels), and "title" fields. Examples: Input → Output: { "poet": "司马光", "poem": "琅菜来从若木边|非膏非沐绿宛延|玉盘委积羞佳客|不是陶家无馔钱", "dynasty": "宋", "sentiments": { "holistic": "implicit positive", "line1": "implicit positive", "line2": "implicit positive", "line3": "implicit positive", "line4": "neutral" }, "title": "昌言有咏石髮诗三章模写精楷殆难复加仆虽未睹兹物而已若识之久者辄复强为三诗以继其后非敢庶几肩差适足为前诗之舆台耳" }. Second example: { "poet": "寇准", "poem": "风露凄清西馆静|悄然怀旧一长叹|海云销尽金波冷|半夜无人独凭栏", "dynasty": "宋", "sentiments": { "holistic": "implicit negative", "line1": "implicit negative", "line2": "implicit negative", "line3": "implicit negative", "line4": "implicit negative" }, "title": "海康西馆有怀" }. Analyze implicit sentiments through semantic density and rhetorical devices. Use temperature 0.3 and top-p 0.9 for annotation consistency. **中文提示词:**你是一个专业分析古典诗歌的专家。请执行FSPC（细粒度情感及副文本分类）任务：1. 通过音形对应模式提取元数据（诗人、朝代、标题） 2. 基于语义语境对齐进行整体与逐行情感分析 3. 输出包含"poet"（诗人）、"poem"（诗歌内容）、"dynasty"（朝代）、"sentiments"（整体及逐行情感标签）和"title"（标题）的JSON格式。示例学习：输入 → 输出：{ "poet": "司马光", "poem": "琅菜来从若木边|非膏非沐绿宛延|玉盘委积羞佳客|不是陶家无馔钱", "dynasty": "宋", "sentiments": { "holistic": "implicit positive", "line1": "implicit positive", "line2": "implicit positive", "line3": "implicit positive", "line4": "neutral" }, "title": "昌言有咏石髮诗三章模写精楷殆难复加仆虽未睹兹物而已若识之久者辄复强为三诗以继其后非敢庶几肩差适足为前诗之舆台耳" }。第二个示例：{ "poet": "寇准", "poem": "风露凄清西馆静|悄然怀旧一长叹|海云销尽金波冷|半夜无人独凭栏", "dynasty": "宋", "sentiments": { "holistic": "implicit negative", "line1": "implicit negative", "line2": "implicit negative", "line3": "implicit negative", "line4": "implicit negative" }, "title": "海康西馆有怀" }。要求通过语义密度与修辞手法识别隐性情感，使用固定参数temperature 0.3和top-p 0.9保证标注一致性，严格保持示例中的JSON字段结构和标签类型。

Figure 13: Prompts of FSPC task within WYWEB dataset: Semantic-form preservation with phonetic-glyph alignment, fixed parameters, strict [semantic-category][format-type] dual-label templates, bilingual instructions enforcing content integrity and structural compliance through classical political/military semantic analysis and folk narrative markers.