Research Square

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

Character Semantic-Phonetic Structure Enhance Language Models in Classical Chinese

Bolin Chang
Nanjing University
Bin Li
lib@njnu.edu.cn
Nanjing Normal University
Zhixing Xu
Nanjing Normal University
Shiyan Ou
Nanjing University

Article

Keywords:

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8179135/v1

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

Additional Declarations: No competing interests reported.

Character Semantic-Phonetic Structure Enhance Language Models in Classical Chinese

Anonymous

Abstract

Writing systems, as fundamental media for human knowledge preservation and facilitating sociocultural interaction, embody the cognitive characteristics of diverse cultures. Despite the rapid advancement of artificial intelligence and language modeling, mainstream language models remain focused on word-level representations, neglecting the rich semantic and structural information embedded within individual characters. In this work, we present an attempt to integrate semantic-phonetic structural information into language modeling, leveraging classical Chinese as the underlying representative system. Our semantic-phonetic-aware language model achieves significant performance gains over the baseline on two core Classical Chinese processing tasks: Word Segmentation and Part-of-Speech Tagging, unveiling the critical role of internal character structure in enhancing language models' representation and encoding capabilities. Through extensive evaluation and in-depth analysis, we reveal that the improvement originates from the synergistic interaction between semantic and phonetic components, as well as the sequential organization of the semantic-phonetic structure. We further argue that achieving the optimal balance between these components is crucial for enhancing the expressiveness and generalizability of language representations. Overall, our findings highlight that the internal organization of characters not only reflects deeper structural principles of the writing system but also opens up a new paradigm for advancing language model design.

1 Introduction

Writing systems serve as primary media for recording languages and convey rich semantic information, representing one of the most fundamental mechanisms of human cognition and sociocultural coordination (Coulmas, 2003; Woods, 2010). Across hundreds of attested writing systems worldwide, they exhibit remarkable diversity and structural complexity (Daniels and Bright, 1996). Logographic systems, such as Egyptian hieroglyphs, primarily organize symbols by semantic cues (Jean, 1992; Wilson, 2004), whereas phonographic systems, including Latin, Greek, and Arabic alphabets, rely on phonological correspondence (Cruttenden, 2021). A third category, exemplified by Chinese characters, integrates both semantic and phonetic characteristics, forming a multidimensional representation system (Handel, 2015). Such structurally-composite writing systems not only enhance visual recognizability and communicative efficiency but also reflect distinct cultural conceptualizations and cognitive patterns shaped through historical evolution and long-term social practice (Watt, 1994; Pae, 2020; Huang et al., 2022).

Contemporary language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b), as well as recent large language models in the artificial intelligence domain, are predominantly trained on text corpora consisting of individual characters. Their preprocessing pipelines rely exclusively on word-level or subword-level tokenization, largely independent of the linguistic typology of the training language (Sennrich et al., 2016; Kudo and Richardson, 2018). These models are then pretrained on self-supervised learning objectives, such as next token prediction, masked language modeling, or next sentence predictions (Rogers et al., 2020). Despite this paradigm has enabled language models to acquire semantically rich and generalizable language representations, it confines and simplifies writing systems to unstructured symbol sequences (Ethayarajh, 2019; Liu et al., 2019a). Consequently, the language characters' structural information is largely omitted, leaving the underlying rich semantic and phonological cues underexplored (Patil et al., 2023). We argue that incorporating character-level structural information into language modeling can substantially enrich the representational capacity of language models and enhance their performance and effectiveness in downstream language understanding tasks.

Traditional Chinese represents a unique and sophisticated writing system that encodes both semantic and phonological information, differentiating it from purely phonographic alphabets (Boodberg, 1937; Boltz, 1994). Hav-

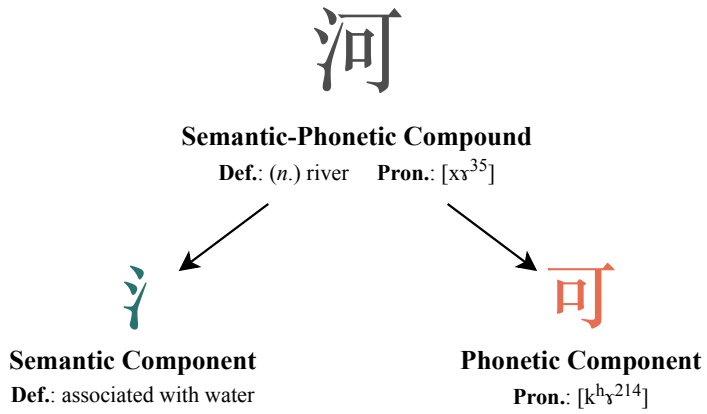


Figure 1: Example of Semantic-Phonetic Compound. *Def.* refers to definition of character or component, and *Pron.* refers to pronunciation in IPA (International Phonetic Alphabet).

ing evolved through millennia, its character structures have been refined and systematized to serve simultaneously as carriers of meaning and as indicators of pronunciation (Boodberg, 1939; Qiu, 2000). Remarkably, over 80% of Chinese characters in modern use are semantic-phonetic compounds (Li et al., 1992), composed of two components: the Semantic Component, which provides categorical cues about meaning, and the Phonetic Component, indicating the character's pronunciation. As illustrated in Figure 1, the character meaning *river* contains a water-related semantic component on the left and a phonetic component on the right, which shares the same vowel as the compound. Some studies about Classical Chinese Processing focus on radicals (Li et al., 2015; Shi et al., 2015; Wu et al., 2021), which constitute only a small part of the character and have limited relevance to its overall linguistic function. In contrast, the semantic-phonetic structure compactly integrates both semantic and phonological information within a single character, offering a well-organized representation that can be systematically integrated into language models. In this work, we adopt Traditional Chinese as a representative writing system to explore the potential of the integration of writing systems' structural properties in enhancing language models' representation learning capabilities.

Following this principle, we designed and developed SemPhoBERT, a Classical Chinese-oriented language model that integrates semantic-phonetic

information from the Classical Chinese writing system. This model is evaluated on two fundamental language modeling tasks in Classical Chinese: Word Segmentation and Part-of-Speech Tagging (POS Tagging). As explicit word boundaries are absent in Classical Chinese literature, identifying word segments serves as the first and fundamental step in Classical Chinese processing (Huang and Zhao, 2007; Zhao et al., 2018; Sun and Zou, 2001). Moreover, POS Tagging identifies grammatical functions and provides essential information for the downstream research on digital humanities (Chiche and Yitagesu, 2022; Zhang and Shang, 2025). Moreover, as both tasks are inherently independent of semantic and phonological cues, they are well-aligned with the dual-encoding mechanism of semantic-phonetic compounds, enabling them to serve as ideal testbeds for evaluating the degree of language representation learning enhancement that our approach could bring to language models (Xue, 2003). Through extensive evaluation, SemPhoBERT demonstrated significant improvement on both tasks compared to the baseline, validating the importance of character-level internal structural information on word boundary detection and lexical categorization tasks, underscoring its broader value for enhancing language models' representation capability.

To conclude, our contributions are summarized as follows. (i) We proposed the framework of explicitly modeling and integrating both semantic and phonetic information into language modeling for semantic-phonetic language systems. (ii) Following this paradigm, we developed SemPhoBERT, a language model capable of fully utilizing the dual-stream information conveyed in Classical Chinese, one of the most representative semantic-phonetic language systems. (iii) Through extensive evaluation, SemPhoBERT demonstrated significantly enhanced performance on both Word Segmentation and POS-Tagging tasks, demonstrating the effectiveness of our framework. (iv) We systematically analyzed the impact of semantic-phonetic structures' interactions with textual representations on language models' performance, and proposed corresponding organizational strategies.

2 Methods

To demonstrate our framework and validate our assumption, we designed a strategy to explicitly model the semantic-phonetic structure and built a semantic-phonetic-aware language model. The constructed pretrained model is then applied to two Classical Chinese downstream tasks: Word Segmen-

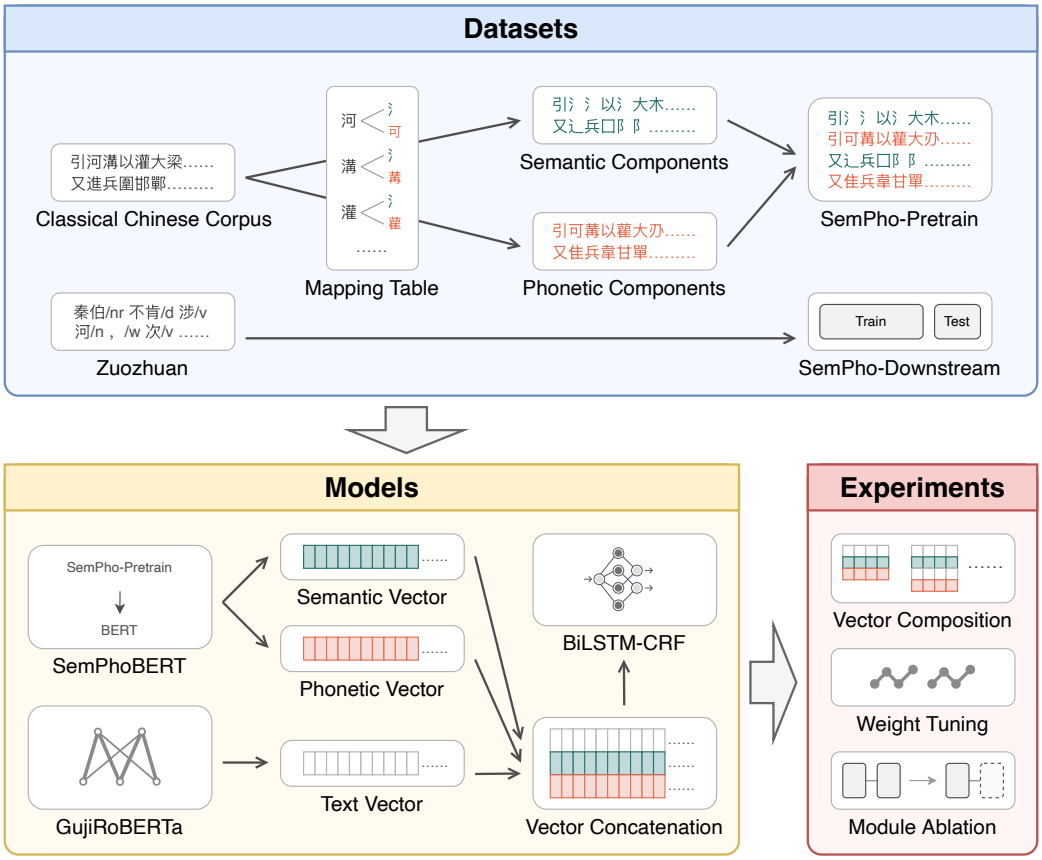


Figure 2: Technical framework of the study, including dataset construction (blue), model architecture (yellow), and experiments (red).

tation and POS Tagging. In this section, we describe our approach in detail from the perspective of 1) dataset construction, 2) model architecture, and 3) experiment designs. We illustrate the technical framework of our work in Figure 2.

2.1 Datasets

We curated and constructed two complementary datasets that serve as the foundation of language model pretraining and downstream-task evaluation. The former, SemPho-Pretrain, contains large-scale, repurposed corpora to model the semantic-phonetic components of Chinese characters explicitly,

aiming at pretraining the model for capturing and utilizing rich semantic-phonetic information embedded in the characters. The latter, SemPho-Downstream, is based on Classical Chinese texts, which are coupled with corresponding word segmentation annotation and POS labels, annotated and verified by human experts. In our experiment, its training split is used for fine-tuning the pretrained language model on two downstream tasks and evaluating the fine-tuned model's performance.

2.1.1 Constructing SemPho-Pretrain

Large-scale language corpora are essential for language models to capture contextual information and latent linguistic patterns during pretraining. To integrate the novel semantic-phonetic structural information, we constructed the pretraining dataset, SemPho-Pretrain, through the repurposing of two resources: (i) a Classical Chinese corpus in Traditional characters and (ii) a character-component mapping table linking characters to their semantic and phonetic components. The dataset is constructed in the following three steps:

Construct Classical Chinese Corpus. *Siku Quanshu* (The Complete Library of the Four Treasuries) is selected as the primary corpus to ensure sufficient linguistic and representation coverage. As the largest known collection of Classical Chinese corpora, it encompasses the four canonical categories: Classics, History, Philosophy, and Literature. Based on the digitized version provided by *Dachengke*[1], we extract the corpus with corresponding punctuations and utilize OpenCC[2] to convert the corpus into Traditional Chinese characters, obtaining a corpus comprising ∼100M characters. The SemPho-Pretrain dataset is constructed based on this foundational corpus.

Character-Component Mapping. We constructed a character-component mapping table for $5,012$ semantic-phonetic compounds in Chinese characters, encompassing each character's simplified and traditional forms, as well as their semantic and phonetic components. Two resources form the basis of our mapping: *Table of General Standard Chinese Characters* (2013), and *Shuowenjiezi* (Xu, 1963). The former was issued in 2013 by the Ministry of

[1]Available at `https://github.com/frankslin/daizhigev20`

[2]Open sourced at `https://github.com/BYVoid/OpenCC`.

Education of the People's Republic of China as an authoritative standard reference for modern Chinese usage, containing 8,105 entries covering both common and sporadic characters. The latter was compiled by *Xu Shen* in the *Eastern Han* dynasty, and is the earliest known systematic dictionary of Chinese characters. It was organized under the framework of the six principles of character formation, providing detailed etymological explanations, which are crucial and valuable for the origin analytics and structural pattern research of semantic-phonetic compounds.

Dataset Construction with Semantic-Phonetic Structure. The Semantic-Phonetic aware pretraining dataset is constructed through mapping curated Classical Chinese corpora into a semantic-phonetic dataset referring to the constructed character-component mapping table. Each semantic-phonetic compound is explicitly decomposed into two separate tokens. The former captures its semantic component while the latter represents its phonetic component, forming a parallel sequence of both semantic and phonetic tokens. Regular characters are duplicated in both sequences to maintain the continuity and coverage of the representation. Both sequences are merged at the document or paragraph level in the end, enabling the joint comprehension of the semantic and phonetic features. Through this process, we constructed a large-scale pretraining dataset featuring explicit semantic-phonetic structure and enriched semantic representations.

2.1.2 Constructing SemPho-DownStream

To evaluate the representation learning capability of our model, we leverage the traditional Chinese dataset *Zuozhuan* (The Commentary of Zuo) from the Linguistic Data Consortium (LDC)[3]. The dataset contains large-scale segmentation and POS label annotations refined and verified by Nanjing Normal University, in accordance with the *Basic Set of Pre-Qin Chinese POS Annotations* in *Information Processing of Pre-Qin Literature* (Chen et al., 2013). It has served as the First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff (Li et al., 2022), organized at the 13th Conference on Language Resources and Evaluation (LREC 2022), allowing our study to be compared with existing works in a fair way.

To repurpose this dataset for fine-tuning and evaluating the SemPhoBERT

[3]Available at `https://catalog.ldc.upenn.edu/LDC2017T14`

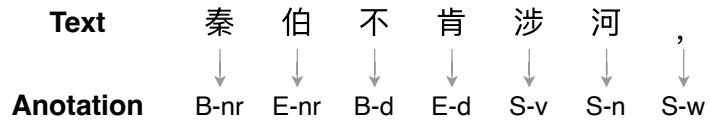


Figure 3: Annotation format for word segmentation and POS tagging. The meaning of this sentence is: "The Duke of *Qin* refused to cross the river."

Subset	#Characters	#Words
Training Set	194,995	166,141
Test Set	33,298	28,131

Table 1: Basic composition of SemPho-Downstream.

base model, we applied a dual-layer annotation scheme. In the segmentation layer, we encode the position of each character within a word using a four-tag scheme B, I, E, S: B denotes the beginning of a multi-character word, I denotes the middle characters, E represents the end of the multi-character word, while single-character words are denoted as S. The POS layer assigns part-of-speech categories to each word following the inventory defined in the *Basic Set of Pre-Qin Chinese POS Annotations*, which includes 21 tags (e.g., v for verb, n for noun). We further combine the annotations in two layers with hyphens (e.g., B-nr) to form a character-level joint representation featuring both the word boundary signal and POS information. Figure 3 demonstrates our annotation format designed for the word segmentation and POS tagging task.

The training and test split of the SemPho-DownStream follows a 6 : 1 ratio. The training split contains the first ten volumes of *Zuozhuan* corpora, while the test split is drawn from its last two volumes. Table 1 demonstrates the composition of SemPho-DownStream.

2.2 Models

In this section, we introduce our semantic-phonetic pre-trained model and the downstream architecture designed to evaluate the contribution of semantic-phonetic structure to language modeling through concrete downstream tasks.

2.2.1 Semantic-Phonetic Pretrained Language Model: SemPhoBERT

To effectively model semantic-phonetic structures, we construct SemPhoBERT by further pre-training BERT on SemPho-Pretrain. BERT has been a widely adopted backbone in Classical Chinese pretrained language models due to its bidirectional Transformer architecture, which supports the learning of contextual semantic and syntactic dependencies. This makes BERT a suitable base for extending the model with semantic-phonetic information. In addition, SemPho-Pretrain provides large-scale sequences of semantic and phonetic components, offering the necessary supervision for integrating these two channels of linguistic information.

In terms of model design, we leverage the inherent correlation between semantic and phonetic components in character formation by pre-training the model on joint semantic-phonetic sequences, rather than training two independent models. This joint-training strategy allows the model to encode each channel's properties while learning their interdependencies within a unified architecture, resulting in more stable and coherent representations across diverse character sets.

In terms of model pre-training, we set the model's hidden size to 768, ensuring sufficient language information representational capacity while maintaining the architectural consistency with GujiRoBERTa [4]. GujiRoBERTa, a RoBERTa-based model pretrained on large-scale Classical Chinese corpora (Wang et al., 2023), is applied in our experiments to embed the text sequence. Our model is trained on SemPho-Pretrain for 5 epochs using AdamW and standard cross-entropy loss as the loss function. The training batch size is set to 3 and the learning rate is fixed to $2e-5$.

2.2.2 Downstream Architecture

For the downstream tasks of Classical Chinese word segmentation and POS tagging, we design a downstream architecture as shown in Figure 4. The architecture consists of two main parts: vector concatenation and the BiLSTM-CRF network.

The processing pipeline starts with converting a Classical Chinese text sequence of length n into two parallel sequences through the character-component mapping table constructed in Section 2.1.1. One sequence con-

[4]Open sourced at `https://huggingface.co/hsc748NLP/GujiRoBERTa_jian_fan`. In this paper, GujiRoBERTa refers to GujiRoBERTa_jian_fan.

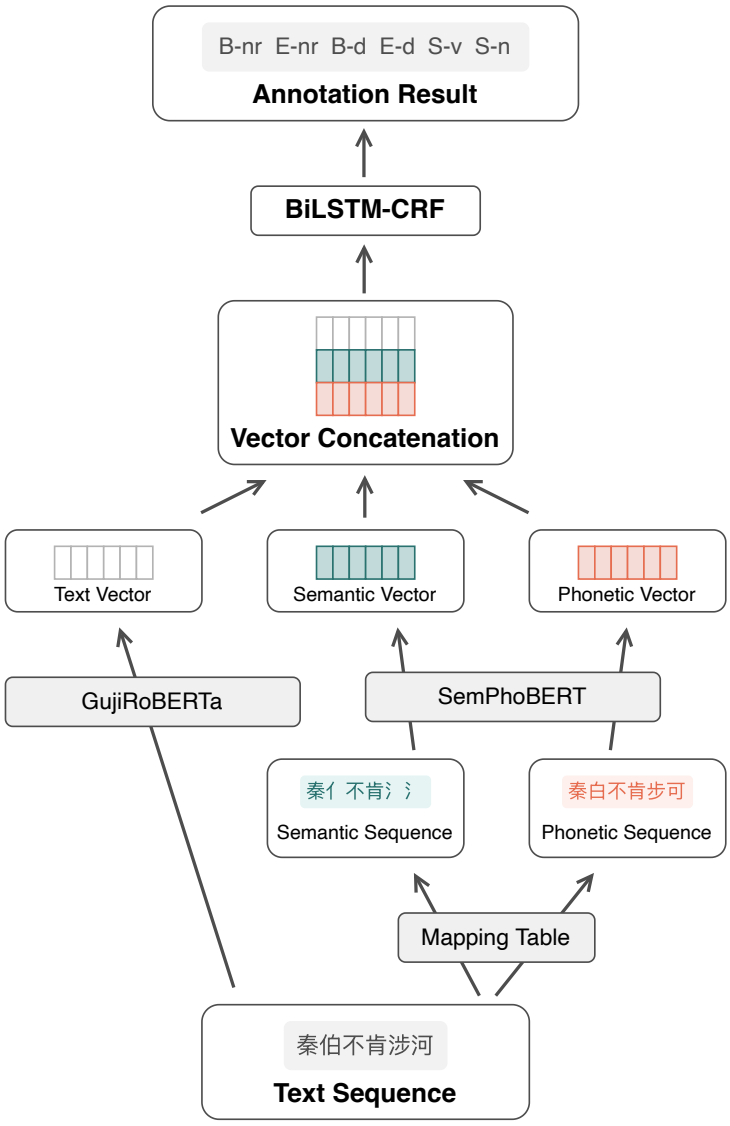


Figure 4: Downstream Architecture.

sists of semantic components and the other consists of phonetic components. These two sequences are then fed separately into SemPhoBERT, which produces two vectors: a semantic vector for the semantic components sequence and a phonetic vector for the phonetic components sequence. Each of these vectors has a dimension of $768 \times n$. In parallel, the text sequence is fed into GujiRoBERTa, which produces a text vector of dimension $768 \times n$.

After obtaining the text, semantic, and phonetic vectors, we concatenate them following the composition described in Section 2.3.1. This concatenation combines textual, semantic, and phonetic information into a single vector whose length remains n, while its width depends on the number of concatenated vectors defined in the chosen composition.

The concatenated vector is then passed into the BiLSTM-CRF network for sequence labeling. BiLSTM-CRF is a widely used framework for sequence labeling (Huang et al., 2015). For this downstream architecture, the BiLSTM layer captures bidirectional contextual dependencies, which are particularly important for Classical Chinese, where word meaning often depends on context (Schuster and Paliwal, 1997). Additionally, the CRF layer models label dependencies explicitly, ensuring global consistency in output sequences (Sutton and McCallum, 2012).

Finally, the output produced by the BiLSTM-CRF network is a sequence of length n, consisting of the predicted word-segmentation markers and POS tags for each character in the input text. These predictions represent the final results of the downstream tasks and reflect the combined contribution of semantic, phonetic, and textual information integrated throughout the model.

2.3 Experiments

To systematically examine the impact of semantic-phonetic structures on language models, we focus on Classical Chinese word segmentation and POS tagging as evaluation tasks and design a series of comparative experiments. The analysis proceeds along three dimensions: vector composition, weight tuning, and module ablation, in order to assess the performance of the semantic-phonetic language model under different conditions and to uncover its underlying patterns. All experiments are conducted on SemPho-Downstream within the downstream architecture.

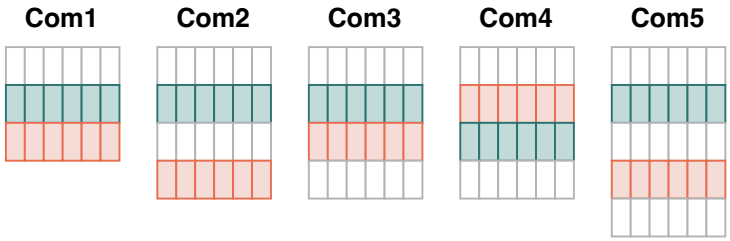


Figure 5: Compositions of vector concatenation. *Com* refers to composition. The top-to-bottom arrangement indicates the concatenation order of the vectors, with the upper ones appended before the lower ones.

2.3.1 Experiment 1: Vector concatenation compositions

To investigate how the organization of vectors influences model performance, we design five concatenation compositions that arrange text vector (T), semantic vector (S), and phonetic vector (P) in different orders and numbers. The five compositions are presented in Figure 5, and their details are described below.

Com1. This composition consists of three vectors: T, S, and P. It is the simplest design, directly concatenating the three vectors in order. Its advantage lies in clarity and low implementation cost, serving as the foundation for more complex compositions. However, with only one text vector, the structure may be insufficient for capturing complex contextual meaning.

Com2. This composition consists of four vectors: T, S, T, and P. Compared with Com1, it inserts an additional text vector between the semantic and phonetic vectors, thereby mitigating the under-representation of textual information. This design provides contextual cues at both the beginning and the middle of the sequence, leading to a more balanced contribution of text vector within the overall compositions.

Com3. This composition also contains four vectors: T, S, P, and T. Relative to Com2, the additional text vector is shifted to the end of the sequence, placing text vector at both boundaries with the semantic and phonetic vectors in between, a configuration that can be viewed as a sandwich arrangement. This design follows the typical structure of Chinese semantic-phonetic compounds.

Com4. This composition contains four vectors: T, P, S, and T. Building on Com3, it reorders the semantic and phonetic vectors, placing the phonetic

vector before the semantic vector. The rationale is to foreground phonetic features, which are more directly tied to pronunciation, thereby enabling the model to better capture phonetic cues.

Com5. This is the most complex composition, consisting of five vectors: T, S, T, P, and T. Relative to Com3, it inserts an additional text vector between the semantic and phonetic vectors. The repeated use of text vector amplifies the role of textual information and promotes deeper feature integration. However, the increased dimensionality also raises computational cost compared with the other composition.

We evaluate all five concatenation compositions to assess their effects. In each case, semantic and phonetic vectors are produced by SemPhoBERT and text vector by GujiRoBERTa, both with a dimensionality of 768, ensuring comparability across semantic, phonetic, and textual features.

2.3.2 Experiment 2: Weight tuning of semantic and phonetic vectors

In semantic-phonetic structures, the semantic component conveys categorical meaning, while the phonetic component provides pronunciation cues. As these two dimensions capture different aspects of internal character information, their roles in downstream tasks may differ. To assess their relative contributions in Classical Chinese word segmentation and POS tagging, we design a set of weighting experiments that compare model performance under different configurations.

Concretely, we assign weights to the semantic and phonetic vectors to control their relative contributions in the model input. Each vector is a $768 \times n$-dimensional representation, and multiplying it by a weight scales its contribution without altering its dimensionality. A weight of 1 retains the original information, while a weight of 0 nullifies it, equivalent to an ablation of that channel in practice. This setup enables the model to capture differences in the relative importance of semantic and phonetic channels during training and inference.

For consistency, we adopt the best-performing concatenation composition from Experiment 1 and design two groups of comparative experiments. In the first, the phonetic vector is fixed at 0.2 while the semantic vector varies from 0.1 to 0.9 (step 0.1); in the second, the semantic vector is fixed at 0.2 while the phonetic vector is varied over the same range. The fixed value of 0.2 is chosen based on preliminary results indicating it as a near-optimal

setting.

These experiments enable a clear comparison of how semantic and phonetic components contribute under different weight settings, shedding light on their respective roles and interactions in downstream tasks.

2.3.3 Experiment 3: Module ablation

To further evaluate the effectiveness of the semantic-phonetic pretrained language model in Classical Chinese word segmentation and POS tagging, we design a set of ablation experiments. The central idea of ablation is to systematically remove different modules within the overall framework to assess their marginal contributions to model performance.

Specifically, we consider four comparative settings: (1) removing both SemPhoBERT and GujiRoBERTa, leaving the BiLSTM-CRF baseline as control; (2) retaining only GujiRoBERTa while removing SemPhoBERT; (3) retaining only SemPhoBERT while removing GujiRoBERTa; and (4) retaining both modules to construct the complete dual-channel model.

All experiments are conducted within the downstream architecture to ensure fair comparison and interpretable results. We adopt the best-performing vector concatenation composition from Experiment 1 and the optimal weight configuration from Experiment 2, ensuring consistency and robustness across settings.

This experiment enables a systematic evaluation of both the individual and combined effects of the text and semantic-phonetic modules, offering a clear understanding of how each functions on its own and how their interaction influences overall performance.

2.3.4 Environment and Parameters

To evaluate the performance of the semantic-phonetic language model on downstream tasks, we conduct a series of experiments on Classical Chinese word segmentation and POS tagging. The hardware and software configurations, together with the key training parameters of the models, are summarized in Tables 2 and 3.

2.3.5 Evaluation

We evaluate model performance on Classical Chinese word segmentation and POS tagging using three standard metrics for sequence labeling: *precision*

Item	Spec
System	Windows 10 Professional
CPU	Intel i7-13700F @ 2.90GHz
GPU	NVIDIA GeForce RTX 4090
CUDA	11.8
PyTorch	1.7.1

Table 2: Hardware and software environment

Hyperparameter	Value
Batch size	32
Hidden size	150
Learning rate	1e-3
Weight decay	5e-5
Optimizer	Adam
Dropout	0.5
Gradient clipping	6.0
Epochs	40

Table 3: Hyperparameter settings

(P), *recall* (R), and *F1-score* (F_1), defined as follows:

$$P = \frac{\text{Correct}}{\text{Predicted}} \times 100\% \tag{1}$$

$$R = \frac{\text{Correct}}{\text{Gold}} \times 100\% \tag{2}$$

$$F_1 = \frac{2PR}{P+R} \times 100\% \tag{3}$$

where:

- *Correct* = number of correctly predicted words with correct POS tags,
- *Predicted* = total number of predicted words,
- *Gold* = total number of gold-standard words in the dataset.

F_1 is the harmonic mean of precision and recall, offering a more balanced and comprehensive assessment of performance. Therefore, we primarily report F_1 as the main evaluation metric.

Evaluation is conducted at the word level rather than the character level, meaning that segmentation and POS tagging are counted as correct only if word boundaries are accurately identified and the whole word receives the correct POS tag.

3 Results

In this section, we present experimental results and analyze how semantic-phonetic structure affects downstream tasks. Their impact on language models depends on two factors: the composition of semantic and phonetic vectors, which shapes representational capacity, and the cross-level interactions that emerge when these features are integrated with textual input, influencing learning and overall performance.

3.1 Effectiveness of Semantic-Phonetic Structures

To evaluate whether semantic-phonetic structures improve model performance, we conducted ablation experiments comparing architectures incorporating SemPhoBERT against the baseline without it. Table 4 reports the results of Experiment 3. As shown in the table, incorporating semantic-phonetic structure substantially enhances performance on both word segmentation and POS tagging in Classical Chinese. Specifically, introducing SemPhoBERT raises the F_1 for word segmentation from 71.55% to 93.08%, an increase of 21.32 percentage points, while for POS tagging the F_1 improves from 49.74% to 85.87%, a gain of 36.13 percentage points. These results demonstrate that semantic-phonetic structure provides significant benefits for Classical Chinese processing by supplying additional structured information to language models.

3.2 Effects of Vector Concatenation Compositions

As a dual encoding mechanism, the semantic and phonetic components of Chinese characters provide complementary semantic and phonological information, and their effectiveness in language models further depends on how

Architecture	Word Segmentation				POS Tagging			
	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1
Base	73.67	69.55	71.55	—	51.22	48.35	49.74	—
Base + Guji	95.93	96.58	96.25	+24.70	91.99	92.60	92.29	+42.55
Base + SemPho	92.75	92.99	92.87	+21.32	85.76	85.98	85.87	+36.13
Base + Guji + SemPho	96.15	96.51	96.33	+24.78	92.22	92.56	92.39	+42.65

Table 4: Results of Experiment 3 (%). *Base* refers to the BiLSTM-CRF baseline used in the ablation experiments. *Guji* refers to GujiRoBERTa, and *SemPho* refers to SemPhoBERT. ΔF_1 indicates the performance gain over the base architecture.

Composition	Word Segmentation			POS Tagging		
	P	R	F_1	P	R	F_1
Com 1: Text + Sem + Pho	95.98	96.54	96.26	91.80	92.33	92.06
Com 2: Text + Sem + Text + Pho	96.20	96.43	96.31	92.05	92.27	92.16
Com 3: Text + Sem + Pho + Text	96.15	96.51	**96.33**	92.22	92.56	**92.39**
Com 4: Text + Pho + Sem + Text	95.92	96.57	96.24	91.88	92.50	92.19
Com 5: Text + Sem + Text + Pho + Text	95.89	96.31	96.10	91.92	92.32	92.12

Table 5: Results of Experiment 1 (%). *Text*, *Sem*, *Pho* respectively refers to text vector, semantic vector and phonetic vector.

they are arranged and preserved. Table 5 reports the results of Experiment 2, which compare different vector concatenation composition. Two major patterns emerge from the analysis.

(1) Continuity between semantic and phonetic vectors improves performance. Compared with Com2, Com3 preserves the same numbers of text, semantic, and phonetic vectors but moves the intervening text vector to the end of the concatenation, thereby maintaining the integrity of the semantic-phonetic information. This adjustment yields small yet consistent improvements, with the F_1 for segmentation increasing by 0.02 and that for POS tagging by 0.23.

Although the improvement is small, it reveals an important principle: the effectiveness of semantic-phonetic structures depends on their arrangement. When the semantic and phonetic vectors are adjacent, they form a coherent representation that strengthens their contribution; when separated by a text vector, this continuity is disrupted and their effect diminishes.

(2) The natural order of semantic and phonetic vectors improves performance. This effect can be observed by comparing Com3 and Com4. Both maintain the same numbers of text, semantic, and phonetic vectors, but

Com3 places the semantic vector before the phonetic vector. This ordering matches the typical structure of semantic-phonetic compounds, in which the semantic component usually comes first. This adjustment yields consistent gains, with the P for segmentation increasing by 0.07 and that for POS tagging by 0.20.

Although the improvement is moderate, it shows that effectiveness depends not only on the coexistence of semantic and phonetic vectors but also on their relative order. When the semantic vector precedes the phonetic vector, the arrangement follows the typical structural pattern of Chinese semantic-phonetic compounds, allowing semantic and phonetic information to integrate more naturally and effectively. If the order is reversed, their contribution diminishes.

3.3 Balancing and Integrating Semantic-Phonetic and Textual Features

The semantic-phonetic structure does not function in isolation; its effectiveness becomes most evident when integrated with textual representations. Comparative analysis of the experimental results reveals two key findings:

(1) Balanced scale between text and semantic-phonetic vectors enhances stability. According to Table 5, Com2, Com3, and Com4 each include two text vectors that match the numbers of semantic and phonetic vectors, whereas Com1, which contains only one text vector, and Com5, which contains three, both show reduced performance. In word segmentation, the P of Com1 (96.26) and Com5 (96.10) are the lowest among the five compositions, falling short of the best result (96.33) by 0.07 and 0.23, respectively. In POS tagging, their scores (92.06 and 92.12) are also lower than the optimal value (92.39), with gaps of 0.33 and 0.27.

These results indicate that model performance is sensitive to the scaling balance between text and semantic-phonetic vectors. Too few text vectors limit contextual support and weaken the effect of semantic-phonetic structure, whereas too many dilute their contribution by lowering their relative scale in the input. Maintaining a balanced numbers between text and semantic-phonetic vectors allows semantic-phonetic structure to occupy a stable role in the input sequence leading to reliable gains in performance

(2) Integration of text and semantic-phonetic vectors maximizes performance. As shown in Table 4, incorporating SemPhoBERT alone

improves P by 21.32 points for segmentation and 36.13 for POS tagging. Incorporating GujiRoBERTa alone yields larger gains of 24.70 and 42.55 points, respectively. When both are incorporated together, segmentation reaches 96.33 (an improvement of 24.78 points over the baseline) and POS tagging reaches 92.39 (an improvement of 42.65 points), the highest scores observed in the experiments.

These results indicate that although semantic-phonetic structure provides clear benefits on their own, their effectiveness is limited without contextual modeling. SephoBERT supplies semantic and phonetic cues but lacks global context, whereas GujiRoBERTa captures contextual and syntactic patterns but ignores character-internal features. When integrated, the two are complementary: contextual modeling is enriched by semantic-phonetic information, allowing all three dimensions to interact within a unified representation space. This synergy delivers the best task performance and fully realizes the value of semantic-phonetic structures in language modeling.

3.4 Balance between Semantic and Phonetic Components

The weight of semantic and phonetic components is a key factor in determining the effectiveness of semantic-phonetic structure in language models. According to the results of Experiment 2 (Figure 6), model performance is evaluated by fixing one weight while varying the other. The outcomes show that performance in both segmentation and POS tagging is highly sensitive to these configurations, underscoring the importance of maintaining the balance between semantic and phonetic components.

(1) Balanced weights yield optimal performance. Performance peaks when both semantic and phonetic weights are set to 0.2, with P of 96.33 for segmentation and 92.39 for POS tagging. This balance enables the two channels to complement each other, as semantic components provide categorical cues while phonetic components provide pronunciation cues.

(2) Imbalance leads to performance degradation. When either semantic or phonetic weight deviates from 0.2, P decline, reaching as low as 96.17 for segmentation and 92.01 for POS tagging in one case, and 96.04 and 91.97 in the other. Such imbalance reflects over-reliance on a single channel, since excessive emphasis on semantic components can obscure broader contextual associations, while excessive emphasis on phonetic components

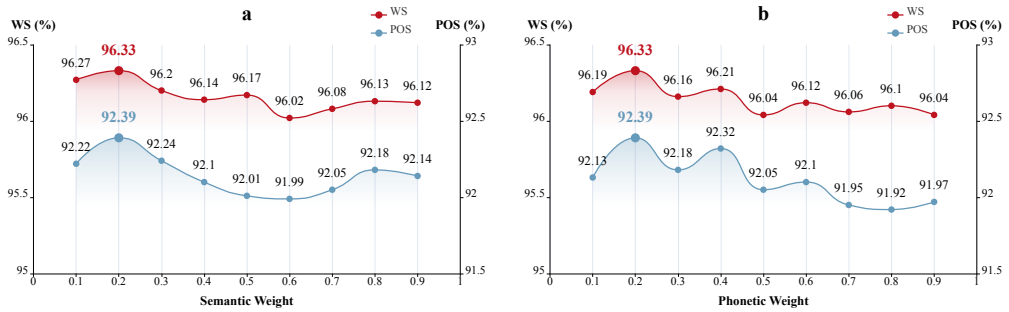


Figure 6: Results of Experiment 2. (a) performance when varying semantic weight with phonetic weight fixed; (b) performance when varying phonetic weight with semantic weight fixed. *WS* refers to the word segmentation task, and *POS* refers to the POS tagging task.

may interfere with morphological and categorical cues, ultimately reducing performance.

In sum, the contributions of semantic and phonetic components are most effective when balanced, as appropriate weighting enables them to complement each other and maximize the impact of semantic-phonetic structure in language modeling.

4 Discussion

This study explores the contribution of internal character structure to language modeling by focusing on the semantic-phonetic compounds of Classical Chinese, which encode both semantic and phonological information. By constructing a semantic-phonetic language model and applying it to word segmentation and POS tagging, we showed that such structure can significantly improve model performance. Our analysis further showed that model performance is strongly influenced by how semantic-phonetic vectors are concatenated, how their weights are balanced, and how effectively these vectors interact with textual representations, highlighting the importance of character-internal structures in language modeling.

Despite these encouraging findings, several aspects remain open for further exploration. First, this study designed five vector concatenation composition, which provide useful insights but do not cover the full range of

possible compositions. Second, the evaluation centered on segmentation and POS tagging, leaving opportunities to extend the analysis to a wider range of downstream tasks. Third, as this study focused on Classical Chinese, extending internal character-structure modeling to other languages and writing systems remains an important direction for future work. These considerations do not detract from the main contributions but rather point to promising avenues for continued investigation.

In conclusion, this study demonstrates that incorporating semantic-phonetic structure can substantially enhance language modeling. Future research should extend this approach to a broader set of Classical Chinese tasks and to other languages in order to assess its cross-linguistic applicability. Such investigations would advance our understanding of character-internal structures in language modeling and offer new directions for multilingual natural language processing.

References

Boltz WG (1994). *The Origin and Early Development of the Chinese Writing System*. American Oriental Society, New Haven.

Boodberg PA (1937) Some proleptical remarks on the evolution of archaic chinese. *Harvard Journal of Asiatic Studies*, 2(2):329–372.

Boodberg PA (1939) 'ideography' or iconolatry? *T'oung Pao*, 35(1):266–288.

Chen X, Feng M, and Xu R (2013) *Information Processing of Pre-Qin Texts*. World Publishing Corporation, Beijing.

Chiche A and Yitagesu B (2022) Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*. https://doi.org/10.1186/s40537-022-00561-y.

Coulmas F (2003) *Writing Systems: An Introduction to their Linguistic Analysis*. Cambridge University Press, Cambridge.

Cruttenden A (2021) *Writing Systems and Phonetics*. Routledge, London.

Daniels PT and Bright W (eds) (1996) *The World's Writing Systems*. Oxford University Press, New York.

Devlin J, Chang MW, Lee K, and Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In Burstein J, Doran C, and Solorio T (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol 1. Association for Computational Linguistics, Minneapolis, p 4171–4186.

Ethayarajh K (2019) How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In Inui K, Jiang J, Ng V, and Wan X (eds) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, p 55–65.

Handel Z (2015) Logography and the classification of writing systems: a response to unger. *Scripta*, 7:109–150.

Huang C, Lin Y, Chen I, and Hsu Y (eds) (2022) *The Cambridge Handbook of Chinese Linguistics*. Cambridge University Press, Cambridge.

Huang C and Zhao H (2007) Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*, 21(3):8–19.

Huang Z, Xu W, and Yu K (2015) Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/1508.01991. Accessed 16 Jun 2025.

Jean G (1992) *Writing: The Story of Alphabets and Scripts*. Harry N. Abrams, New York.

Kudo T and Richardson J (2018) Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco E and Lu W (eds) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, p 66–71.

Li B, Yuan Y, Lu J, Feng M, Xu C, Qu W, and Wang D (2022) The first international ancient chinese word segmentation and pos tagging bakeoff: Overview of the evahan 2022 evaluation campaign. In Sprugnoli R and Passarotti M (eds) *Proceedings of the Second Workshop on Language*

Technologies for Historical and Ancient Languages. European Language Resources Association, Marseille, p 135–140.

Li Y, Kang J, Wei L, and Zhang S (1992) Research on modern chinese semantic-phonetic characters. *Applied Linguistics*, 74–83.

Li Y, Li W, Sun F, and Li S (2015) Component-enhanced chinese character embeddings. In Màrquez L, Callison-Burch C, and Su J (eds), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, p 829–834.

Liu NF, Gardner M, Belinkov Y, Peters ME, and Smith NA (2019a) Linguistic knowledge and transferability of contextual representations. In Burstein J, Doran C, and Solorio T (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol 1. Association for Computational Linguistics, Minneapolis, p 1073–1094.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, and Stoyanov V (2019b) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692. Accessed 16 Jun 2025.

Ministry of Education of the People's Republic of China (2013) Table of general standard chinese characters. Beijing.

Pae HK (2020) *Script Effects as the Hidden Drive of the Mind, Cognition, and Culture*. Springer, Cham.

Patil R, Boit S, Gudivada V, and Nandigam J (2023) A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Qiu X (2000) *Chinese Writing*. University of California, Berkeley.

Rogers A, Kovaleva O, and Rumshisky A (2020) A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Schuster M and Paliwal K (1997) Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Sennrich R, Haddow B, and Birch A (2016) Neural machine translation of rare words with subword units. In Erk K and Smith NA (eds) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol 1. Association for Computational Linguistics, Berlin, p 1715–1725.

Shi X, Zhai J, Yang X, Xie Z, and Liu C (2015) Radical embedding: Delving deeper to Chinese radicals. In Zong C and Strube M (eds) *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol 2. Association for Computational Linguistics, Beijing, p 594–598

Sun M and Zou J (2001) A critical appraisal of the research on chinese word segmentation. *Contemporary Linguistics*, 3(1):22–32, 77.

Sutton C and McCallum A (2012) An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Wang, D, Liu C, Zhao Z, Shen S, Liu L, Li B, Hu H, Wu M, Lin L,Zhao X, and Wang X (2023) Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. arXiv preprint arXiv:2307.05354. https://arxiv.org/abs/2307.05354. Accessed 16 Jan 2025.

Watt WC (ed) (1994) *Writing Systems and Cognition*. Springer, Dordrecht.

Wilson P (2004) *Hieroglyphs: A Very Short Introduction*. Oxford University Press, Oxford.

Woods C (ed) (2010) *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. The Oriental Institute of the University of Chicago, Chicago.

Wu S, Song X, and Feng Z (2021) MECT: Multi-metadata embedding based cross-transformer for chinese named entity recognition. In Zong C, Xia F, Li W, and Navigli R (eds) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol 1. Association for Computational Linguistics, Online, p 1529–1539.

Xu S (1963) *Shuowenjiezi*. Zhonghua Book Company, Beijing.

Xue N (2003) Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*. p 29–48.

Zhang H and Shang J (2025) Chinese Word Segmentation and Part-of-Speech Tagging. In *Natural Language Processing and Applications*. Springer Singapore, Singapore, p 173–203.

Zhao H, Cai D, Huang C, and Jie C (2018) Chinese Word Segmewtation: Another Decade Review (2007-2017). In *Frontiers of Empirical and Corpus Linguistics*. China Social Science Press, Beijing, p 139–162.