Automatic Reconstruction of Ancient Chinese Pronunciations

Zhige Huang*, **Haoan Jin***, **Mengyue Wu**†, **Kenny Q. Zhu**†
Shanghai Jiao Tong University, Shanghai, China
University of Texas at Arlington, USA
{yvonne_huang, pilgrim, mengyuewu}@sjtu.edu.cn,
kenny.zhu@uta.edu

Abstract

Reconstructing ancient Chinese pronunciation is a challenging task due to the scarcity of phonetic records. Different from historical linguistics' comparative approaches, we reformulate this problem into a temporal prediction task with masked language models, digitizing existing phonology rules into ACP (Ancient Chinese Phonology) dataset of 70,943 entries for 17,001 Chinese characters. Utilizing this dataset and Chinese character glyph information, our transformer-based model demonstrates superior performance on a series of reconstruction tasks, with or without prior phonological knowledge on the target historical period. Our work significantly advances the digitization and computational reconstruction of ancient Chinese phonology, providing a more complete and temporally contextualized resource for computational linguistics and historical research. The dataset and model training code are publicly available[1].

1 Introduction

A human language is comprised of a pronunciation system and a writing system, both evolving and changing over time. An important part of historical phonology is about reconstructing the historical pronunciation system and its pattern of evolution, sometimes as drastic as the three examples in Table 1[2]. Studies on Chinese historical phonology using modern comparative methods began in 1915 (Karlgren, 1915), followed by various attempts to reconstruct MiddleTang Chinese (Karlgren, 1922; Sagart, 1991; Wang, 2012). These studies reveal the causes and trends of language evolution and contribute to the study of ancient literature resources.

Char	T	L	S	Y	Q	M
顺	[dʑiuin]	[ʑiuən]	[ɕiuən]	[ɕyən]	[ʂun]	[ʂuən]
席	[ziɐk]	[ziək]	[sit]	[si]	[si]	[ɕi]
乏	[biuæp]	[viuæp]	[fap]	[fau]	[fa]	[fa]

Table 1: Reconstructed pronunciations (in IPA) of three characters over 6 Chinese historical periods: MiddleTang - T, LateTang - L, Song - S, Yuan - Y, MingQing - Q and Modern - M. See Figure 3 for corresponding timeline.

Historical scripts are often preserved in tangible forms, yet the transmission of pronunciation through successive generations is more susceptible to variation and distortion. Across linguistic domains, the systematic evolution of phonetics frequently eludes precise documentation. In the realm of Sinological phonology inquiry, even though the characters' glyph information can provide some insight into phonetic attributes, their pronunciations cannot be deterministically inferred from their orthographic form. (See Figure 1). Therefore, unlike phonograms in many other common languages, Chinese characters are considered logographic, meaning each phoneme or syllable does not correspond to a specific character.

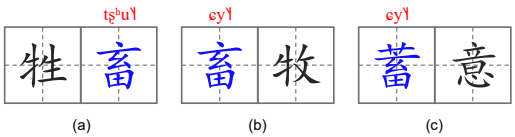

Figure 1: (a) and (b) show that one character may have different pronunciation in different words; (b) and (c) show different characters with same glyph component may share same pronunciation.

When attempting to reconstruct the Chinese phonological system for a certain historical period, only limited resources of pronunciation records are provided. Current studies are based on a rhyming

* These authors contributed equally to this work.
† Corresponding authors.
[1] https://github.com/KaguraRuri/Ancient-Chinese-Phonology
[2] Pronunciations in this paper are all transcribed with International Phonetic Alphabet (IPA). See https://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/ for IPA characters and sounds.

dictionary *Guangyun* (Chen, 1936) published in AD 601. From then on, Chinese linguists in different dynasties inherited the taxonomy of character pronunciations in *Guangyun*, and used the comparative method to group similar-sounding characters into the same category. Combined with xeno-Chinese transliteration[3] , modern dialect pronunciation, etc., the Chinese pronunciation system can be partially reconstructed (Zhao, 2015).

Given ample attempts and combining different linguistic discoveries, the rule-based reconstruction method has its inherent limitations. First, the pronunciation is reconstructed within each category but not for each character, making the result neither intuitive nor easy to search. Second, current reconstruction results cannot cover all Chinese characters' pronunciation over all historical periods. Chinese language system's evolution is not strictly under the taxonomy of *Guangyun* (See Sec. 2), making it impossible to always apply rule-based change patterns on a whole category[4]. Third, current results are not fully digitized but only in printed versions. As a matter of fact, the process of character-wise pronunciation reconstruction is complex, encompassing the entirety of Chinese characters across various historical eras, and necessitates prior knowledge in linguistics, thereby calling for further refinement.

Recently, several computer-assisted methods have been proposed to reconstruct ancient Chinese pronunciation, including Chang et al. (2022)'s comparative Chinese dialect dataset and Kim et al. (2023)'s approach of applying transformer model on this dataset to rebuild ancient Chinese pronunciation. However, the reconstruction results are limited to a specific historical period, i.e., Middle-Tang Chinese pronunciation. Hence, we attempt to build a time-aware dataset and use a temporal factor-embedded model to complete the reconstruction at an arbitrary time point.

The contribution of this paper is as follows:

- We build a chronological ACP (**A**ncient **C**hinese **P**ronunciation) dataset by combining and digitizing the *Guangyun*-based Chinese character taxonomy and the existing ancient Chinese reconstruction results in linguistics, offering 70,943 pronunciation entries for 17,001 Chinese characters (Sec. 2).
- Aiming to interpolate and extrapolate the reconstruction result to any time point, we propose a transformer-based pronunciation reconstruction model (Sec. 3). With additional language features encoded, our model achieves the best accuracy score on random-split, phonetic distinction, and reduced training data evaluations compared to baseline models (Sec. 4), showing its ability to refine incomplete phonological reconstruction results of traditional linguistics.
- By using a chronological dataset, our time-aware model also has the ability to reconstruct the pronunciation for a given period when the information for training is sparse or even completely missing in the current ACP dataset.

2 ACP Dataset Construction

Our **ACP** (Ancient Chinese Pronunciation) dataset offers character-wise chronological data of ancient Chinese pronunciation, combining two kinds of data: the digitized *Guangyun* data (KanjiDatabase-Project, 2004) and the phonological reconstruction result of Wang (2012). For a given character, the former informs us the category it belongs to, and the latter tells us the pronunciation reconstruction results on each category.

Guangyun is a rhyme dictionary using a special sound annotation called *Fanqie*. Chinese characters are monosyllabic, i.e. all Chinese character's pronunciation correspond to one syllable, each comprised of an initial and a final (Duanmu, 2007). According to *Fanqie*, each Chinese character's pronunciation is described as a combination of two representative characters, one for its *initial* and another for its *final* as shown in Figure 2.

Figure 2: Method of *Fanqie*: "东" belongs to category "德"([t]) for initial and to category "红"([uŋ])for final.

Under this taxonomy, there are 38 categories of initials and 298 categories of finals. The aim of phonological reconstruction is to attach the exact pronunciation (denoted by IPA phoneme) onto

[3]Translation materials between Chinese and foreign languages in a specific period, e.g. Translation of Sutras published in MiddleTang can correspond Sanskrit (phonographic language) to MiddleTang Chinese (logographic language).

[4]For the reconstruction of the intermediate dynasties (Yuan, Ming and Qing dynasties), the linguistic reconstruction we use can only cover some of the Chinese characters for which rhyming patterns can be hypothesized based on the available documentary materials.

each initial and final category for targeted period. These information can be found in Wang (2012)'s reconstruction results of Chinese phonology, which includes the evolution of Chinese pronunciation system from PreQin (-206 BC) to modern Chinese (AD 1912-) at 9 different representative time points. We have only selected results for 6 historical periods from MiddleTang dynasty and beyond: `MiddleTang`, `LateTang`, `Song`, `Yuan`, `MingQing` and `Modern`, represented by the beginning each historical period (AD 581, AD 836, AD 960, AD 1279, AD 1368, AD 1912), since these reconstruction results are more consistent among different linguists (Tang, 2011; Wang, 2012). See Figure 3 for the Chinese history timeline.

When constructing ancient Chinese pronunciation for each character in ACP dataset, 4 possible cases are presented. Concrete examples for different cases of reconstruction can be found in Table 2.

Direct determination of pronunciation If the exact pronunciation is directly given for one category, then the given IPA phoneme is attached onto each character within this category. For example, "波" belongs to initial category of "幫" and [p] is attached to category "幫", then the initial IPA of "波" is [p]. This is the case for pronunciation system of MiddleTang since all characters strictly belongs to its category denoted in *Guangyun*.

Rule-based determination of pronunciation If there exists several possible pronunciations for the same category, then the linguistic rules given by Wang are applied to help choose the correct one. Rules on initials' pronunciation are usually based on finals' category, and rules on finals' pronunciation are usually based on the articulation information recorded in *Guangyun* [5]. For example, "跗"="幫"+"废" and "碑"="幫"+"支" both belongs to initial category of "幫", but given the linguistic rule that "幫" represents [f] only under the case when final category is 废, and represents [p] in other cases, we attach [f] as 跗's initial IPA and [p] as 碑's initial IPA. This is the case for LateTang and Song period since a small amount of categories' pronunciation has encountered rule-based change.

Arbitrary determination of pronunciation If a category-wise pronunciation reconstruction is no given due to the complexity of the language system's evolution, then we manually digitize Wang's example on several representative characters' pronunciation. This is the case for Yuan and MingQing pronunciation system since the overall categorization has significantly changed compared to MiddleTang phonology system.

Converted pronunciation For modern Chinese language system, we take Mandarin as its representative. Since Mandarin is a living language, we directly convert the pronunciation to IPA representation.

Following these IPA symbols and conditions, we manually attach the initials and finals, take all the available records as single entries and left the unknown entries blank. Meanwhile, another articulation information, the tone, is much more complicated to trace since rhyme dictionaries focus on initials' and finals' taxonomy and only give vague description on tone (Wang, 2012). The linguistic reconstructions have no consistent results (Yuchi, 1986; Shi, 1983; Pan, 1982), thus neither encoded into ACP dataset for any of the time period.

According to the articulation feature, we further divide the final part into Medial, Nucleus and Coda as shown in Figure 4. The medial is an optional part of the final, usually has short and soft sounding, connecting initial and final. The nucleus is the main and non-empty vowel in final. The coda is attached to nucleus which can only be consonant or empty. Finally, each single entry in our dataset is composed of 4 parts: Initial-I, Medial-M, Nucleus-N and Coda-C, either empty (denoted by "-") or non-empty (denoted by one IPA phoneme). Example of complete and incomplete sets of pronunciations for one character in the dataset can be found in Table 3. The final dataset contains 17,001 entries for each of the 17,001 Chinese characters in MiddleTang, LateTang, Song and Modern historical period. Meanwhile, only 1,402 entries for Yuan and 1,519 entries for MingQing is provided while other filled with [UNK] due to the inherent incompleteness in linguistic reconstruction (see Table 4).

3 Pronunciation Reconstruction Model

In this section, we introduce the architecture of our pronunciation reconstruction model, centered around Figure 5. By utilizing an embedding layer and a transformer encoder layer, our model is capable of learning complex patterns and relationships within both the phonetic data and the glyph infor-

[5] The articulation information is only vaguely recorded in *Guangyun*.

Case	Character	Initial	Final	Rule	Initial IPA
direct	波	帮	戈	帮=[p] and 戈=[uɑ]	[p]
rule-based	砩	帮	废	if(initial=帮 and final=废) then [f] else [p]	[f]
	碑	帮	支		[p]
arbitrary	方	帮	阳	-	[f]
converted	比	帮	旨	-	[p]

Table 2: Five different examples of reconstruction in four different cases constructing our ancient Chinese pronunciation dataset for each category. For an identical initial category, different rules applied can lead to different reconstruction result for initial IPA.

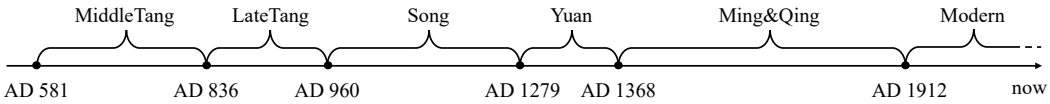

Figure 3: China history timeline. The 6 pronunciation systems represents the overall spoken language form in a certain period of time.

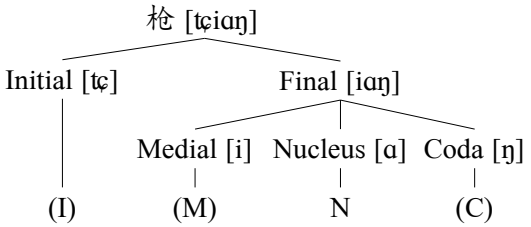

Figure 4: The phoneme sequence is split into initial, medial, nucleus and coda, where initial, medial and coda are optional.

mation, enabling accurate reconstructions across different historical periods.

3.1 Model Architecture

As shown in Figure 5, our glyph and temporal enhanced (**GTenhanced**) pronunciation reconstruction model consists of an embedding layer and a transformer encoder layer.

First, in order to capture the phonetic information delivered by Chinese character glyph, we build upon the approach of Lyu et al. (2021). Specifically, using the Han Ideographs structure files[6], we follow the methodology outlined by Ke and Hagiwara (2017)[7] to generate glyph trees for all Chinese characters in our dataset. Each glyph tree is then converted into a sequence format using a depth-first algorithm, as described by Nguyen et al. (2019). As shown in Figure 6, the input sequence comprises two distinct token types: leaf nodes (positions 3, 4, 5) representing components of Chinese characters, and internal nodes (positions 1, 2) representing structural operators (e.g., left-right). These token types will be utilized for glyph type embedding. The positions of these tokens (0, 1, 2, 3, 4, 5) denote the sequence order in the input and will subsequently be utilized for glyph position embedding, as shown in Figure 5.

Next, for the feature representation of Chinese character pronunciations, we segment the syllable of a Chinese character into four parts, as mentioned in Sec. 2. The entire input sequence consists of the character glyph sequence, the temporal sequence, and the sequences of phonetic changes over time for each part of the syllable. For example, the final sequence representation for the character "暂" can be serialized as shown in Figure 5.

3.2 Embedding Layer

The model input embedding is the combination of `token, type, position, and character segmentation` embeddings, as shown in Figure 5.
Token Embedding encodes glyph, temporal and phoneme features in different segments started by token [CHAR], [YEAR] and [IPA] and followed by glyph, temporal and phoneme information. Additionally we use [MASK] to mask the phoneme tokens, use [UNK] to fill in the unknown Yuan and MingQing phoneme data.
Type Embedding distinguishes token types: for glyph feature tokens, "CHAR" denotes the character tag [CHAR], "STC" and "CPN" represents the structure and component type; for temporal features, "YEAR" marks the period tag [YEAR] and "NUM" signifies the numerically encoded year information; for phoneme tokens, "IPA" labels the phoneme tag [IPA], while "PHO" designates the

[6] https://github.com/tomcumming/chise-ids
[7] https://github.com/yuanzhiKe/Radical_CR_Encoder

Character	MiddleTang				LateTang				Song				Yuan				MingQing				Modern			
	I	M	N	C	I	M	N	C	I	M	N	C	I	M	N	C	I	M	N	C	I	M	N	C
纡	d	i	ou	-	d̥	i	əu	-	tɕ	i	əu	-	tɕ	i	əu	-	ʂ	-	əu	-	tʂ	-	ou	-
雍	ʔ	i	u	ŋ	ʔ	i	u	ŋ	j	i	u	ŋ	UNK				j	-	y	ŋ	j	-	u	ŋ
晢	s	-	i	k	s	i	ə	k	s	-	i	t	UNK				UNK				ɕ	-	i	-

Table 3: Example of complete and incomplete sets in the dataset.

	T	L	S	Y	Q	M
Entries	17,001	17,001	17,001	1,420	1,519	17,001
Coverage	100%	100%	100%	8.25%	8.93%	100%

Table 4: The statistics of ACP dataset. Abbreviations: T - MiddleTang, L - LateTang, S - Song, Y - Yuan, Q - MingQing, M - Modern.

phoneme.

Position Embedding assigns a number starting from 0 to each token within the same feature, helping the model understand token positions in their respective sequences.

Segmentation Embedding identifies different features, helping the model distinguish between character glyph, historical periods, and phonetic changes. All the embedding have the same dimension d.

3.3 Masked Transformer Encoder

We utilize the multi-head self-attention network as the foundational structure. Given a sequence of tokens represented by $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n denotes the number of tokens in the sequence and d represents the dimension of each token, the process of masked self-attention can be formulated as follows:

$$\mathbf{A} = \frac{(\mathbf{X}\mathbf{W}^Q)(\mathbf{X}\mathbf{W}^K)^\top}{\sqrt{d_k}}$$
$$\widetilde{\mathbf{X}} = \mathrm{Softmax}(\mathbf{A} + \mathbf{M})(\mathbf{X}\mathbf{W}^V),$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_k}$ serve as learnable parameters, while $\mathbf{M} \in \mathbb{R}^{n \times n}$ denotes the attention mask matrix (Liu et al., 2020). We derive $\mathbf{M}$ by setting $\mathbf{M}_{ij}$ to 0 when x_j is visible to x_i, and to $-\infty$ when x_j is invisible to x_i. Specifically, all tokens within the same feature are mutually visible to each other; additionally, the special tags [CHAR], [YEAR], and [IPA] are also mutually visible to each other. The Softmax function is applied to the attention scores to normalize them into a probability distribution, ensuring that the weights sum to one.

3.4 Training Target

Inspired by the training methodology of BERT (Devlin et al., 2018), we randomly mask 15% of the phoneme tokens in the input sequence. Within this 15%, 80% are replaced by the mask token [MASK], 10% are randomly replaced by a token belonging to the same token type, and 10% remain unchanged. Consequently, the model is trained to predict the original phoneme tokens based on the modified input, as illustrated in the top-right of Figure 5(b).

4 Evaluations

In this section, we comprehensively evaluate the performance of GTenhanced Transformer across various evaluation tasks in our dataset.

4.1 Experimental setup

In this subsection, we provide an overview of 5 tasks designed to test various aspects of the model's reconstructive capabilities compared to several baseline models.

4.1.1 Baselines

We compare our model to four baseline models. The random daughter and majority constituent method are from Chang et al. (2022) but we use an improved version. For each part of the syllable (Initial, Medial, Nucleus and Coda), a random phoneme (random daughter) or a most frequently appearing phoneme (majority constituent) is chosen from inputs of each available historical period and then combined into a syllable as reconstruction result. For decision tree classifier, the reconstruction is also done on each of the four parts. We also adapted cognate transformer (Akavarapu and Bhattacharya, 2023), which utilizes both row and column attention to reconstruct the phoneme on each position. Since this model was designed for proto-word reconstruction task where all inputs are contemporary pronunciations, time factor can be embedded but meaningless for our chronological language reconstruction.

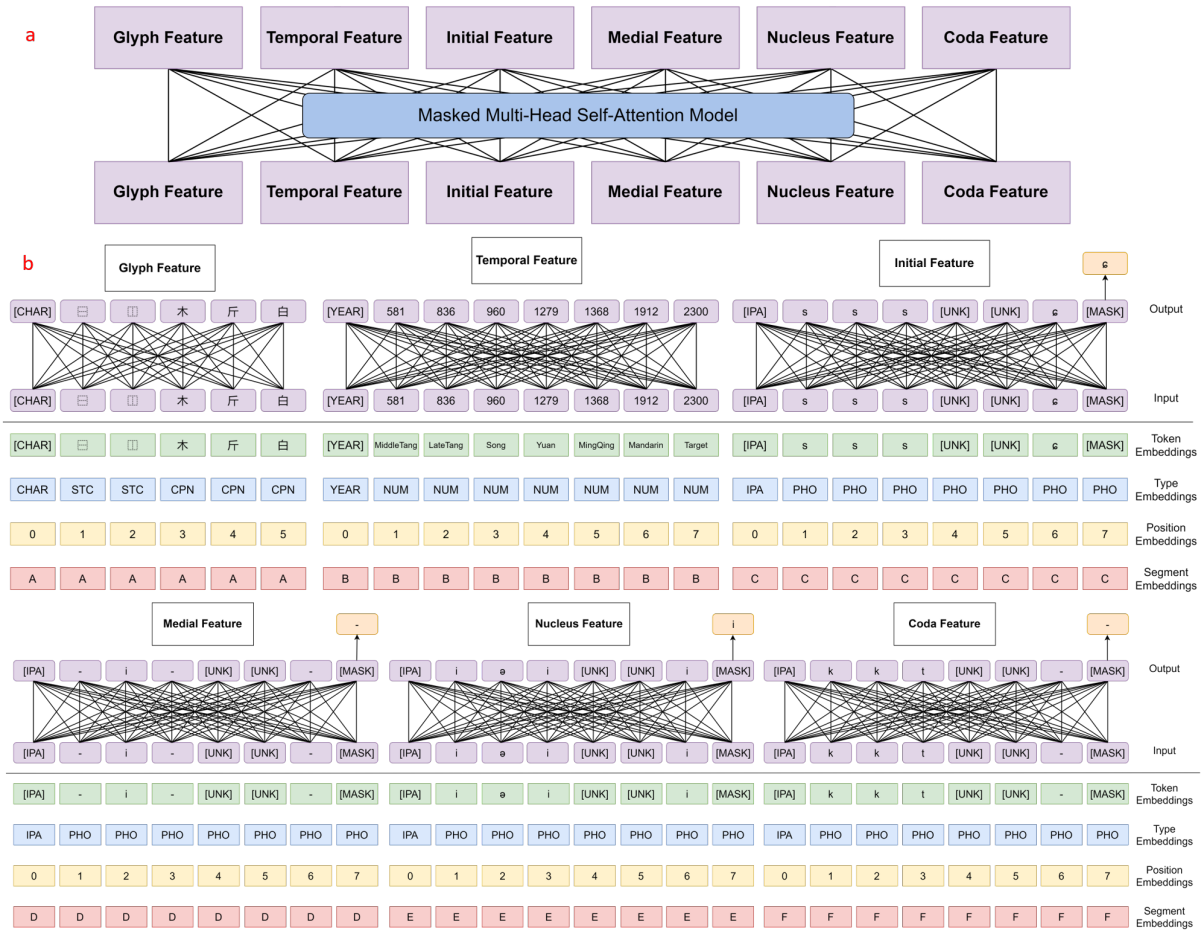

Figure 5: **Architecture of glyph and temporal enhanced ancient Chinese pronunciation reconstruction model. (a)** Using a feature-differentiated block architecture, the model transmits attention between blocks through special markers such as [CHAR], [YEAR], and [IPA]. **(b)** The embedding of glyph feature, temporal feature, initial feature, medial feature, nucleus feature, and coda feature.

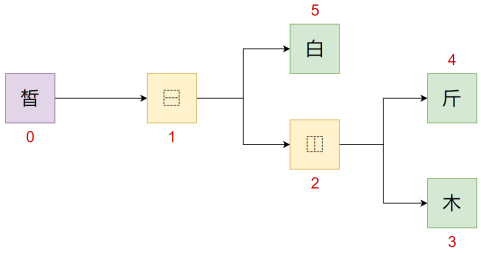

Figure 6: The glyph tree of a Chinese character.

4.1.2 Evaluation Tasks

In this section, we describe the tasks designed to evaluate the performance of GTenhanced Transformer, each testing different aspects of its reconstructive capabilities.

Random Split Evaluation The dataset is randomly split into training and testing sets with a 7:3 ratio. Due to substantial incomplete data for the Yuan and MingQing periods, we first partition the dataset into four subsets: characters missing both Yuan and MingQing pronunciations, characters missing only Yuan pronunciations, characters missing only MingQing pronunciations, and characters with no missing data. Each subset is then split into training and testing sets using the same seed for randomization, ensuring a 7:3 ratio. The subsets are then combined to form the final training and testing datasets.

Phonetic Distinction Evaluation Characters with phonetically same Modern pronunciations are segregated to ensure they do not appear in both the training and testing sets, increasing the difficulty of the task. The dataset is first divided into four subsets as in the Random Split Evaluation, then split into training and testing sets while maintaining phonetic distinction, and finally combined to form the final datasets.

Evaluation with Reduced Training Data from the Reconstructed Era This task involves decreasing the amount of training data from the reconstructed era. For example, to reconstruct Modern pronunciations, the training set may contain only a fraction of the available Modern data or none at all. The training and testing sets are split as in the Phonetic Distinction Evaluation, ensuring no overlap of phonetically similar characters between sets.

Evaluation with Reduced Historical Training Data We progressively reduce the historical phonetic data available for training to assess the model's performance under varying levels of data scarcity. For example, to reconstruct Modern pronunciations, we provide data from only the MiddleTang, LateTang, and Song periods, or fewer. The training and testing sets are split as in the Phonetic Distinction Evaluation.

Predict Future Pronunciation This task predicts possible future pronunciations using the known pronunciations from six historical periods: MiddleTang, LateTang, Song, Yuan, MingQing, and Modern. The model's predictions are purely speculative due to the absence of ground truth data. This exploration offers insights into the model's capacity for extrapolation and generalization beyond historical contexts.

4.2 Experiment Results

Random Split Evaluation Table 5 shows our model's superior performance in reconstructing pronunciations across all historical periods in the random split task. The results shown are averaged over three runs. Despite significant data gaps in the Yuan and MingQing periods, our model consistently achieves an F1 score above 0.85. In contrast, the decision tree model's performance suffers due to extensive missing data during these periods, highlighting our model's robustness in handling incomplete datasets.

Furthermore, compared to the Cognate Transformer model, our approach exhibits a slight advantage in reconstructing pronunciations for the Yuan and MingQing periods. This edge is attributed to our model's ability to effectively integrate glyph and temporal features, enabling a nuanced understanding of phonetic evolution over time and facilitating accurate reconstructions in data-sparse periods.

Model	T	L	S	Y	Q	M
RD	0.167	0.179	0.181	0.157	0.166	0.155
MC	0.175	0.179	0.196	0.194	0.207	0.219
DT	0.947	0.976	0.953	0.442	0.353	0.787
CT	0.958	0.965	0.923	0.810	0.838	0.867
GTeT	**0.961**	**0.980**	**0.972**	**0.852**	**0.873**	**0.876**

Table 5: Model performance on random split evaluation (Metrics: F1). Abbreviations: RD - Random Daughter, MC - Majority Constituent, DT - Decision Tree, CT - Cognate Transformer, GTeT - GTenhanced Transformer.

Phonetic Distinction Evaluation Table 6 shows that our model still maintains optimal performance and a high F1 score even under the strict partitioning of the training and testing sets. The results are also averaged over three runs. In this scenario, characters with the same pronunciation do not appear in both the training and testing sets simultaneously. However, by leveraging glyph and temporal features, our model can accurately reconstruct target pronunciations from related phonetic information. This demonstrates the model's ability to generalize and infer pronunciations based on learned patterns, even when direct phonetic similarities are not present in the training data.

Model	T	L	S	Y	Q	M
RD	0.167	0.179	0.181	0.157	0.166	0.155
MC	0.175	0.179	0.196	0.194	0.207	0.219
DT	0.821	0.889	0.794	0.131	0.171	0.451
CT	0.863	0.928	0.855	0.613	0.574	0.500
GTeT	**0.931**	**0.942**	**0.933**	**0.702**	**0.652**	**0.728**

Table 6: Model performance on phonetic distinction evaluation (Metrics: F1).

Evaluation with Reduced Training Data from the Reconstructed Era Figure 7 and Table 7 depict the findings from our evaluation with reduced training data from the reconstructed era. Here, the decision tree model's performance diminishes linearly as training data decreases. In contrast, attention-based models like the Cognate Transformer and our GTenhanced Transformer exhibit a logarithmic decline in performance under reduced training conditions, indicating their resilience to data reduction.

Our GTenhanced Transformer notably maintains a significant F1 score even when no training data for M pronunciations is available. This resilience stems from its ability to leverage character glyph and temporal features, facilitating accurate reconstructions based on related historical data. These

results underscore the robustness of our model in handling sparse datasets, highlighting its practical potential where complete data is often lacking.

As shown in Table 7, both the Decision Tree and Cognate Transformer models exhibit zero performance (F1 score of 0) when there is no training data from the reconstructed era. The Decision Tree model relies on patterns seen during training to make reconstructions, rendering it ineffective without target-era data. Similarly, the Cognate Transformer model's use of row and column attention fails without target-era training, hindering its ability to establish meaningful connections for accurate reconstructions across historical periods.

Moreover, the decline in F1 scores with the reduction of target period data in the training set further validates the effectiveness of our dataset. The dataset's richness in historical and phonetic context is crucial for accurate pronunciation reconstruction, and the model's performance drop with less data underscores this importance.

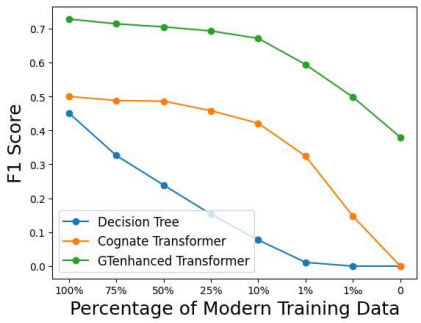

Figure 7: Model performance on evaluation with reduced training data from the reconstructed era.

Evaluation with Reduced Historical Training Data As shown in Figure 8, we progressively reduce the historical context data when reconstructing Modern pronunciation. The F1 score decreases more slowly compared to reconstructing MiddleTang pronunciation. Specifically, when reconstructing Modern pronunciation, the F1 score drops from 0.380 to 0.285 as we reduce the available historical context from T+L+S+Y+Q to only T. On the other hand, when reconstructing MiddleTang pronunciation, the F1 score drops from 0.682 to 0.283 as we reduce the historical context from L+S+Y+Q+M to only M. The F1 scores become nearly identical at the final stages. This phenomenon stems from the model's heavier reliance on phonetic information and attention weights from adjacent eras, particularly MiddleTang, LateTang, and Song periods, which exhibit structured and rule-based phonetic patterns.

Additionally, the decline in F1 scores also validates the effectiveness of our dataset. As we reduce the historical context, the model's performance drops, indicating that the available historical phonetic information is crucial for accurate pronunciation reconstruction.

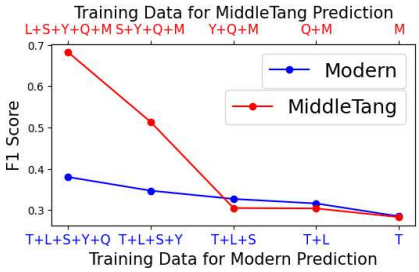

Figure 8: Model performance on evaluation with reduced historical training data.

Predict Future Pronunciations Our model has demonstrated robust performance, maintaining a certain level of F1 score even in the absence of training data for specific historical periods. To further explore the capabilities of our model, we conducted an intriguing experiment to predict the pronunciation of Chinese characters in AD 2300[8]

5 Related Work

Language models for phonetic reconstruction A related task of phonological ancient language reconstruction is proto-word reconstruction, which takes set of words in different contemporary languages as input and the corresponding word in their common ancestral language as result of supervised reconstruction. Meloni et al. (2021) and Akavarapu and Bhattacharya (2023) both evaluated neural networks' performance on Romance language family's reconstruction task. Kim et al. (2023) first introduced Transformer architecture into proto-word reconstruction task and outperforms previous models on both Romance and Sinitic dataset. While large language models (LLMs) have recently demonstrated exceptional capabilities in understanding and generating contemporary languages, their proficiency in comprehending ancient Chinese, remains inadequate. Zhang and Li (2023)'s research highlighted the limitations

[8] You can listen to the audio representations of future Chinese pronunciations at: https://github.com/KaguraRuri/Ancient-Chinese-Phonology.

Model	100%	75%	50%	25%	10%	1%	1‰	0
Decision Tree	0.451	0.326	0.239	0.153	0.077	0.011	0	0
Cognate Transformer	0.500	0.488	0.486	0.458	0.421	0.324	0.147	0
GTenhanced Transformer	**0.728**	**0.714**	**0.705**	**0.693**	**0.671**	**0.594**	**0.498**	**0.380**

Table 7: Model performance on Reduced Target Training Data Evaluation (Metrics: F1). The target of the model reconstruction is Modern pronunciation. The value of the header represent the percentages of Modern pronunciation data in the training set relative to the entire training set. The division between the training and testing sets follows the phonetic distinction evaluation.

of LLMs in handling the complex ancient Chinese phonetic information.

Chinese phonetic dataset In terms of Chinese phonetic datasets, current digitization all organized the ancestor language (Middle Tang Chinese) and its daughter languages (modern Chinese dialects) into a cognate set. Hou (2004) first collected 2,789 cognates of word-wise Chinese dialect pronunciation. Chang et al. (2022) expanded Hou's dataset, organize entries by characters instead of word. As for chronological phonology dataset in Chinese, existing resources are mainly from studies of historical linguistics. Swedish sinologist Karlgren first put forward the phonological reconstruction of Middle Tang Chinese (Karlgren, 1922). Wang (2012) provided a comprehensive analysis of Chinese language phonological evolution. However, these sources are not digitized to our knowledge.

6 Conclusion

We introduce an extensive ancient Chinese pronunciation dataset with 70,943 entries for 17,001 Chinese characters, alongside an enhanced transformer-based model integrating glyph and temporal information to refine traditional phonological reconstruction results. Our model outperforms traditional methods across various ancient Chinese pronunciation reconstruction tasks with superior accuracy even under low-resource scenarios. Despite the incomplete phonetic data, it maintains high performance for reconstructing and predicting Chinese pronunciations. We offer a richer, temporally contextualized resource for computational linguistics and historical research. This study lays a strong foundation for future research in phonetic reconstruction and language evolution.

Limitations

Despite the significant advancements made by our approach, several limitations remain. First, our current dataset does not encode tone information and its evolution, which may be beneficial crucial for accurately reconstructing ancient Chinese pronunciation and for educational purpose. Future work will focus on enhancing the dataset by incorporating detailed tonal information. Furthermore, the dataset currently lacks non-linguistic features such as geographical, natural, and political factors that could influence phonetic changes over time. Including these features could provide a more holistic understanding of ancient Chinese phonetic reconstruction and improve the model's accuracy, especially on low-resource scenarios.

Addressing these limitations would enhance the robustness and applicability of our methodology, thereby advancing the field of computational reconstruction of ancient Chinese phonology.

References

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023. Cognate transformer for automated phonological reconstruction and cognate reflex prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6852–6862. Association for Computational Linguistics.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A New Comparative Dataset for Chinese Languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3563–3569. International Committee on Computational Linguistics.

Pengnian Chen. 1936. *廣韻: 5卷 [Guangyun: 5 volumes]*. 中华书局[Zhonghua Book Company].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

San Duanmu. 2007. *The phonology of standard Chinese*. OUP Oxford.

Jingyi Hou. 2004. *Xiandai Hanyu Fangyan Yinku* 现代汉语方言音库 *[Phonological database of Chinese dialects]*. 上海教育出版社[Shanghai Educational Publishing House].

KanjiDatabaseProject. 2004. 宋本廣韻データ [Songben Guangyun].

Bernhard Karlgren. 1915. *Études sur la phonologie chinoise*. Leyde et Stockholm.

Bernhard Karlgren. 1922. The reconstruction of ancient chinese. *T'oung Pao*, 21(1):1 – 42.

Yuanzhi Ke and Masafumi Hagiwara. 2017. Radical-level ideograph encoder for rnn-based sentiment analysis of chinese and japanese. In *Asian Conference on Machine Learning*, pages 561–573. PMLR.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed protoform reconstruction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–38. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-Bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

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.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473. Association for Computational Linguistics.

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.

Wuyun Pan. 1982. 关于汉语声调发展的几个问题 [Several problems in the development of Chinese tones]. 中国语言学报 *[Journal of Chinese Linguistics]*, 10(02):359–386.

Laurent Sagart. 1991. Lexicon of reconstructed pronunciation in early middle chinese, late middle chinese and early mandarin vancouver: University of british columbia press. *Cahiers de Linguistique Asie Orientale*, 20(2):247 – 248.

Xiangdong Shi. 1983. 玄奘译著中的梵汉对音和唐初中原方音 [Sanskrit-Chinese syntax in Xuanzang's translations and pronunciation in the Central Plains of the Early Tang dynasty]. 语言研究*[Studies in Language and Linguistics]*, pages 27–48.

Zuofan Tang. 2011. 汉语语音史教程 *[Hanyu Yuyinshi Jiaocheng]*. 北京大学出版社[Peking University Press].

Li Wang. 2012. 汉语语音史 *[A History of Chinese Phonetics]*. 中华书局[Zhonghua Book Company].

Zhiping Yuchi. 1986. 日本悉昙家所传古汉语调值 [pitch of tones for ancient chinese transmitted by japanese siddhanta scholars]. 语言研究*[Studies in Language and Linguistics]*, (02):17–35.

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

Tong Zhao. 2015. 汉语音韵学概论 *[Introduction to Chinese Historical Phonology]*. 中国人民大学出版社 [China Renmin University Press].