BIRD: Bronze Inscription Restoration and Dating

Wenjie Hua
Wuhan University
huawenjie@whu.edu.cn

Hoang H. Nguyen
University of Illinois, Chicago
hnguy7@uic.edu

Gangyan Ge
Xinjiang University
gegangyan@163.com

Abstract

Bronze inscriptions from early China are often fragmentary and chronologically uncertain. We introduce **BIRD** (**B**ronze **I**nscription **R**estoration and **D**ating), a fully encoded dataset with standard scholarly transcriptions and chronological labels. We further propose an allograph-aware masked language modeling framework combining domain- and task-adaptive pretraining with a Glyph Net (GN), which links graphemes and allographs. Experiments show that GN improves restoration, while glyph-biased sampling enhances dating.

1 Introduction

Bronze inscriptions from the Chinese Bronze Age (21st–3rd c. BCE) are among the most important early Chinese textual sources (Li, 2024). Found on ritual vessels, weapons, and musical instruments, these inscriptions record military achievements, feudal enfeoffments, oaths, and ancestral rites. Yet as excavated texts, they are often fragmentary and damaged, with uncertain chronological assignments.

Traditional restoration and dating rely on expert comparison of parallel expressions and contextual inference, along with other features, a process that is labor-intensive and difficult to scale. Neural models, particularly pre-trained language models (PLMs), have recently shown promise in supporting ancient text restoration. However, existing applications of artificial intelligence to bronze inscriptions focus almost exclusively on computer vision, such as single-character recognition or denoising of inscription images (Guo, 2021; Zhao et al., 2020). By contrast, natural language processing (NLP) approaches to inscriptional texts remain largely unexplored, despite their potential for tasks such as restoration and dating.

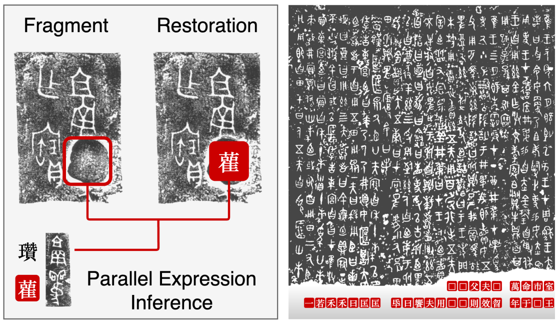

Fragment
Restoration
瓚
雚
雚
Parallel Expression Inference
□□父夫□ 萬命市室
一若禾禾日匡匡 卑日饗夫用□□則效曶 年于□王

Figure 1: Left: A simplified paleographer's workflow for restoring a damaged bronze inscription: identifying the damaged fragment, inferring from parallel expressions, and proposing a restoration (Zeng, 2011; Wu, 2012; Xie, 2014). Right: A damaged bronze inscription fragment (CCYZBI.02838A) (CASS, 2007) with the expert's inferred reading (Huang, 2022). The workflow mirrors a masked language modeling setup, where restorations are hypothesized from local context and attested parallel expressions.

Two factors make NLP modeling of bronze inscriptions challenging (Li, 2024):

1. **Low-resource setting.** Although nearly 20k inscriptions have been published, most are extremely short, with over half containing three or fewer characters. Compared to modern corpora, the effective training data is therefore sparse.
2. **Allography.**[1] In the Western Zhou corpus alone, 2,134 graphemes include 572 allograph sets (48.15%) (Liu, 2009). Current encodings treat such forms as separate tokens, which prevents semantically equivalent allographs from being learned as a unified grapheme, thereby hindering generalization in data-hungry Transformers. Figure 2 shows a representative family of allo-

wjhuah/BIRD

[1] We use the term *allograph* for distinct graphical forms that realize the same grapheme, following the graphematic perspective of Meletis (2020, 2019). In Chinese palaeography, these correspond to so-called *yitizi*. As Qiu (2013) notes, the broad definition of *yitizi* subsumes two subtypes: narrow allographs (fully interchangeable forms) and partial allographs (forms that once overlapped in usage but later diverged, functionally close to *tongyongzi*).

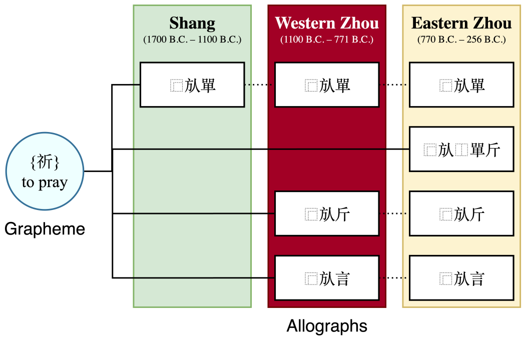


Figure 2: Concrete glyph family of *Qi* ('to pray') from the Shang to the Eastern Zhou. To illustrate the correlation between glyphs and their components, Ideographic Description Sequences (IDS) are used.

graphs that share the same grapheme.

Nonetheless, bronze inscriptions belong to the same synchronic register as transmitted and excavated Pre-Qin (21st–3rd c. BCE) texts (Li, 2024), which can be leveraged as auxiliary data for domain-adaptive pretraining (Gururangan et al., 2020). Moreover, allographic variation is not mere noise: in downstream tasks, e.g., chronological dating, glyph distinctions provide evidence (Wang, 2015; Su, 2016). Effective modeling thus requires careful use of normalization to support low-resource learning, while still retaining the distinctive historical signal.

The contributions of this paper are as follows:

- BIRD: the first fully encoded bronze inscription dataset (41k tokens) with encoding suitable for NLP tasks;
- The construction of Glyph Net (GN), a resource that pairs and clusters graphemes and their allographs into sets;
- A framework for allograph-aware modeling that integrates GN for restoration and glyph-biased sampling for dating.

2 Related Work

Research on bronze inscriptions has a long history. Paleographic resources, notably the widely accepted compilation of CASS (2007), form the basis of BIRD, complemented by the translations and philological studies of Wu (2012), Ma (1986), and Shirakawa (1962). Chronological issues have also been examined, with frameworks proposed by Tang (2016), Chen (2004), and Guo (1999).

Digitization efforts have made significant contributions. The *Digital Retrieval Platform for Shang and Zhou Bronze Inscriptions (Jihewang)* platform[2] integrates catalogs, glyph images, lexica, etc. Academia Sinica has released two semi-open databases: the *Digital Archives of Bronze Images and Inscriptions (AS DABII)*[3], covering vessel images, rubbings, typology, and metadata; and the *Lexicon of Pre-Qin Oracle, Bronze Inscriptions and Bamboo Scripts (AS Lexicon)*[4], spanning oracle bones, bronzes, and bamboo manuscripts for lexical research. However, these resources remain ill-suited for NLP tasks, as many characters, especially allographs, are represented only as images. Hence, addressing allography is crucial, with studies focusing on collecting examples across dynasties (Qi, 2023; Du, 2020; Su, 2016; Luo, 2013).

Neural model restoration of fragmentary texts has been well-explored across languages. Most related to our work, Mo et al. (2021) applied BERT (Devlin et al., 2019) to masked character prediction on the Shanghai Museum bamboo manuscripts (1–9, 2,103 characters), simulating the speech case induction. Wang et al. (2025) further combined RoBERTa (Liu et al., 2019) with computer vision for restoring incomplete Chinese steles. In other low-resource epigraphic domains, similar approaches have achieved strong performance on Latin inscriptions (Assael et al., 2025), Arabic manuscripts (Miloud et al., 2024), Greek inscriptions (Assael et al., 2022), and Akkadian cuneiform (Lazar et al., 2021). Chronological dating tasks have also been pursued (Assael et al., 2025; Chen et al., 2024; Tian and Kübler, 2021).

Distinct from prior work, we provide the fully encoded and chronologically labeled bronze inscription corpus, accompanied by a grapheme-allograph resource, which enables neural models to tackle both restoration and dating.

3 Dataset

3.1 Pre-Qin Corpus (DAPT)

We perform domain-adaptive pretraining on Pre-Qin texts, covering 40 works across 11 categories with a total of 2.09M tokens, which were compiled from open corpora including the *Chinese Text Project*[5] and Wikisource[6], and were further normalized (Appendix B).

[2] https://jwdcdbz.ancientbooks.cn

[3] https://bronze.asdc.sinica.edu.tw

[4] https://inscription.asdc.sinica.edu.tw

[5] https://ctext.org

[6] https://wikisource.org

Dataset	Ava.	Dedup.	Filt.	Enc.	Chron.
Jihewang	✗	✗	✗	Partial	✓
AS DABII	✗	✗	✗	Partial	✓
AS Lexicon	✗	✗	✗	Partial	✓
BIRD	✓	✓	✓	Full	✓

Table 1: Comparison of bronze inscription digitization efforts. Our dataset is the only publicly available, deduplicated, and filtered corpus, with complete encoding and chronological labels.

3.2 BIRD (TAPT)

Existing resources for bronze inscriptions, such as *Jihewang*, *AS DABII*, and *AS Lexicon*, primarily serve as retrieval platforms rather than structured datasets. To fill this gap, we introduce BIRD, the first NLP-ready dataset for inscription restoration and dating.

BIRD comprises 41k tokens and incorporates paleographic and digitization scholarship (Section 2). Each inscription is annotated with the dynasty and finer-grained period, which supports supervised experiments on chronological classification. A comparison with previous digitization efforts is shown in Table 1.

In addition, we provide a Glyph Net resource of 1,078 grapheme–allograph pairs, compiled from Shang, Western Zhou, and Eastern Zhou studies (Qi, 2023; Du, 2020; Luo, 2013). Following the principle of mutual substitutability (Qiu, 2012), it supports glyph family–level reasoning for MLM and downstream tasks.

BIRD is prepared through four steps:

1. **Encoding.** All inscriptions are converted into machine-readable text, with characters categorized into three types: identifiable, unreadable (□), and undeciphered (`[UNK]`; see Appendix F), as summarized in Table 2.
2. **Filtering.** Extremely short inscriptions (≤1 character; 6,078 out of 17,547 in *AS DABII*), mostly redundant single-character marks (e.g., "Shi Ding" consisting only of the character "Shi," CCYZBI.01073–01088 (CASS, 2007)), are removed to avoid trivial patterns and ensure a more representative corpus.
3. **Deduplication.** Many inscriptions recur across vessels (e.g., ten identical "Bo Xian Fu Li," CCYZBI.00649–00658 (CASS, 2007)), as exact formulaic repetitions. Keeping all copies would inflate token counts, cause overfitting to duplicated patterns, and risk leakage, so we retain a single representative instance.

Type	Count	Proportion
Identifiable	39,565	99.24%
Unreadable (□)	236	0.59%
Undeciphered (`[UNK]`)	56	0.14%

Table 2: Types of tokens and their proportions in BIRD.

4. **Correction.** Clerical transcriptions (*liding*) and chronological assignments are revised according to recent philological studies, with efforts made to align the corpus with the most up-to-date interpretations. Further details are provided in Appendix G.

4 Model

We use standard Transformer (Vaswani et al., 2017) masked-language-model (MLM) backbones, which have proven effective in text restoration tasks. Applying it to bronze inscriptions, however, presents two challenges: (i) the low-resource nature of the corpus, and (ii) the prevalence of allographs, where semantically equivalent forms appear as distinct tokens.

To address these issues, we introduce three modifications to the MLM pipeline: (1) domain-adaptive pretraining (DAPT) on a contemporaneous Pre-Qin corpus, with shallow layers frozen to stabilize training; (2) Glyph Net (GN), constructed from grapheme–allograph pairs. Transitive closure induces glyph families, and newly observed glyphs are aligned to family centroids; (3) a glyph-biased sampling strategy that leverages GN families in two ways: aligning allographs for restoration, and emphasizing historically informative allographs for dating, inspired by weighted sampling techniques shown to enhance MLM training (Zhang et al., 2023). Figure 3 illustrates the overall architecture.

5 Experiments

5.1 Baselines

We evaluate a BiLSTM sequence model as the restoration baseline (Luong et al., 2015; Sutskever et al., 2014), and an SVM classifier, which has shown strong performance in dynasty classification of historical Chinese texts (Tian and Kübler, 2021). For pretrained backbones, we consider MULTILINGUALBERT (mBERT), XLM-ROBERTA (base and large) (Conneau et al., 2020), and SIKUROBERTA (Wang et al., 2021). Multilingual mBERT and XLM-R have demonstrated strong

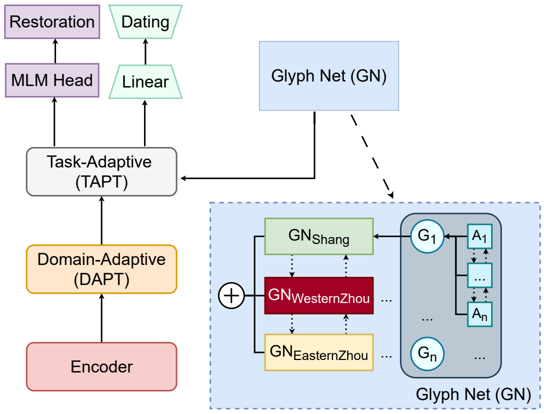


Figure 3: Our pipeline enhances masked language modeling for bronze inscriptions by combining domain-adaptive pretraining (DAPT), task-adaptive pretraining (TAPT), and Glyph Net module (as illustrated in the lower-right component, each grapheme $G_{1..n}$ is linked to its allographs $A_{1..n}$) that integrates allograph glyph information into a BERT or RoBERTa backbone.

transfer performance in low-resource and cross-lingual settings (Lazar et al., 2021; Chau et al., 2020), while domain-specific models trained on the *Siku Quanshu* corpus are widely adopted in ancient Chinese NLP (Hua and Xu, 2025; Ge, 2022; Mo et al., 2021).

5.2 Implementation Details

Bronze inscriptions are extremely short, so standard BERT-style random masking often removes nearly all available context. To mitigate this, we adopt a stride-based masking scheme (s) that masks every s-th non-boundary character, so that sequences of length $\leq s$ lose at most one token. The stride parameter is tuned for each backbone using Bayesian hyperparameter search with Weights & Biases (Biewald, 2020).

5.3 Tasks

We model two complementary tasks that reflect real palaeographic challenges. For **restoration**, we apply the stride-based masking scheme (Section 5.2), and require the model to recover the gold character from incomplete inscriptions. Predictions are evaluated both at the exact character level and at the glyph-family level, where allographs under the same grapheme are treated as interchangeable.

For **dating**, we fine-tune a linear head on the encoder representations to predict both dynasty-level and finer-grained period labels. The same backbone, settings, and adaptation schedules are shared across both tasks, which ensures comparability.

Model	Params	E@1	E@5	E@10	F@1	F@5	F@10
BiLSTM	20M	39.02	42.98	53.10	**57.41**	57.63	62.50
SikuRoBERTa	109M	**49.47**	**65.20**	**70.15**	54.32	**68.05**	**73.07**
mBERT	110M	43.55	58.57	63.71	46.93	61.28	65.92
XLM-Base	278M	43.51	58.35	62.94	44.28	59.49	64.03
XLM-Large	550M	45.64	60.92	64.91	47.16	61.17	65.36

Table 3: Restoration performance comparison of backbone models under the unified **GN** setting. **E@K** = Exact match at rank K; **F@K** = Family-level match at rank K. All scores are percentages.

6 Results and Discussion

6.1 Evaluation Criteria

For **restoration**, we follow prior work (Assael et al., 2022; Lazar et al., 2021) in single-position prediction. Performance is measured using: *Exact@K*, which checks if the gold token appears within the top-K predictions, and *Family@K*, which counts a prediction correct if any member of the gold token's allograph family appears within the top-K. All results are evaluated on glyph forms unseen during training, to approximate real restoration scenarios.

For **dating**, we evaluate at two granularities: dynasty-level (Shang, Western Zhou, Spring and Autumn, Warring States period), and period-level (Early, Middle, Late). We report accuracy and macro-F1, and additionally compute a hierarchical score that first verifies the dynasty label and then the period label within the predicted dynasty.

6.2 Restoration Results

Table 3 shows restoration results. SIKUROBERTA achieves the best performance on five of six metrics, including 49.47 Exact@1 and 73.07 Family@10, outperforming BiLSTM by +10.5 p.p. (Exact@1) and +10.6 p.p. (Family@10). BiLSTM leads on Family@1 (57.41). Multilingual PLMs lag SIKUROBERTA by 4–6 p.p. on Exact@1, which confirms the advantage of in-domain pretraining.

6.3 Dating Results

Table 4 reports dynasty- and period-level dating. SIKUROBERTA delivers the best overall performance (dynasty accuracy 86.42; macro-F1 77.83) and the highest hierarchical dynasty accuracy (84.21). MBERT trails by 1–4 points, while larger multilingual encoders are less competitive. XLM-BASE benefits slightly from period distinctions, and XLM-LARGE surpasses SIKUROBERTA on Hier-Per F1 but lags elsewhere. Overall, period dating proves more challenging than dynasty dating.

Model	Params	Dynasty		Hier-Dyn		Hier-Per	
		Acc	F1	Acc	F1	Acc	F1
SVM	0.08M	75.31	49.44	76.32	42.67	58.55	49.43
SikuRoBERTa	109M	**86.42**	**77.83**	**84.21**	**54.32**	**67.11**	62.91
mBERT	110M	84.57	74.77	82.24	53.12	63.82	58.63
XLM-Base	278M	79.01	50.34	80.92	51.32	62.50	57.34
XLM-Large	550M	84.01	74.60	81.58	53.12	65.13	**62.96**

Table 4: Dating performance comparison of backbones under the unified **glyph-biased** sampling. **Dynasty** = four-way classification; **Hier-Dyn / Hier-Per** = hierarchical evaluation at dynasty and period levels. All scores are percentages.

6.4 Analysis

For **restoration**, GN is the most consistent contributor. On the SikuRoBERTa backbone, GN achieves the highest Exact@K and Family@K (Table 9), with an average of 58.3 across metrics, surpassing Bias (57.8) and no adaptation (56.5), which demonstrates that collapsing allographs into glyph families effectively reduces sparsity and stabilizes restoration.

For **dating**, the trend reverses. GN alone contributes little; instead, glyph-biased sampling toward glyph tokens yields the best dynasty- and period-level results (Table 10), averaging 68.9 across metrics—2.4 and 1.3 p.p. higher than GN (66.5) and GN+Bias (67.6), respectively. This empirically confirms that glyph distinctions serve as robust diachronic markers (Wang, 2015; Su, 2016), while combining GN with bias adds no further gain.

Takeaway: restoration benefits from modeling allographic equivalence, whereas dating exploits diachronic differentiation.

6.5 Error Patterns

Restoration is strongest in formulaic segments with fixed patterns (Ma, 2003), while nouns denoting vessels, temporal adverbs, and modal particles achieve high accuracy due to their syntactic stability (Wu, 2023). Errors mainly default to frequent templates and confusions among semantically related nouns, verbs, or numerals, though even mispredictions often preserve syntactic category awareness.

Misclassifications in dating concentrate in the Spring and Autumn and Warring States periods, where inscriptions are stylistically freer (Ma, 2003) and dynasty boundaries blur. Class imbalance also skews errors toward the Western Zhou. Nevertheless, severe cross-era errors remain rare, which indicates that the models effectively capture chronological signals.

7 Conclusion

We present BIRD (**B**ronze **I**nscription **R**estoration and **D**ating), a curated dataset of transcribed and chronologically labeled bronze inscriptions. On top of this dataset, we design a masked language modeling framework that integrates domain-adaptive pretraining (DAPT), task-adaptive pretraining (TAPT), and allograph-aware training. Our experiments show that this framework, especially with the SikuRoBERTa backbone, achieves state-of-the-art results in both restoration and dating: Glyph Net (GN) stabilizes restoration, while glyph-biased sampling enhances chronological classification. We hope BIRD provides a foundation for future NLP research on Chinese bronze inscriptions.

Limitations

Despite promising gains in both restoration and dating, several limitations remain. First, BIRD-still suffers from sparsity and long-tail imbalance, which constrains generalization for rare forms. Related effort in this area can be found in (Li et al., 2025; Nguyen et al., 2020) and similar studies.

Second, glyph-level modeling remains a challenge. Different characters may not consistently represent the same word (Qiu, 2013), and our Glyph Net currently relies merely on inductive bias at the family level. Its generalization could be strengthened by incorporating stricter palaeographic constraints (Chou and Huang, 2005) and expanding the knowledge base of loan characters (Wang et al., 2023). Moreover, diachronic distributions of allographs are not well modeled, and the system is prone to semantically plausible but orthographically inappropriate predictions.

Third, our setup lacks phonological supervision. Bronze and other early Chinese inscriptions frequently employ loans (Baxter and Sagart, 2014), yet sound-based substitution is invisible to a token-only model. Incorporating phonetic series embeddings may capture such regularities, following phoneme-aware strategies that have proven effective in non-Latin scripts (Nguyen et al., 2025, 2023).

Fourth, fragmentary evidence poses a major obstacle. Some inscriptions are partially legible, with subcomponents visible even when the full graph is nearly damaged. A token-level MLM cannot leverage such partial signals. Structure-aware encodings, such as Ideographic Description Sequences (IDS), have been shown to be effective in related tasks (Yu et al., 2023; Pan et al., 2026), which could enable component-conditioned modeling for more robust restoration and dating.

Fifth, our framework omits archaeologically multimodal signals that are central to traditional dating. The shape of the vessel, the decorative motifs, and the casting techniques provide independent chronological evidence (Chen, 2004), but are not yet explored. Integrating textual modeling with such modalities would bring the system closer to expert chronological practice.

Finally, our experiments are constrained by computational resources. We primarily relied on BERT- and RoBERTa-based backbones for sequence modeling due to their efficiency, whereas more recent generative architectures may better capture long-range dependencies and support free-form restoration. Exploring such models remains an open direction. Consequently, this work is positioned as a standardized dataset and a baseline framework to facilitate future research. Predictions should be viewed as auxiliary hypotheses, which offer preliminary guidance to paleographers, with expert interpretation and archaeological context remaining indispensable.

Ethics Statement

This work relies exclusively on ancient Chinese texts and bronze inscriptions, which contain no personal or sensitive information. The models are intended solely for academic research, and their predictions should not be regarded as authoritative readings of the inscriptions. We also acknowledge the environmental impact of model training, though our experiments involve relatively small-scale models with limited computational cost.

Acknowledgements

We would like to thank Dan Liu, Bin Li, Yuwen Zhang, Youzu He, Rongwei Yi, and the anonymous reviewers for their valuable feedback during the iterations of this paper. This project was supported in part by Wuhan University under the project “Research on Pre-Qin Inscriptions and Early Official Documents” (No. S202510486008).

References

Yannis Assael, Thea Sommerschield, Alison Cooley, Brendan Shillingford, John Pavlopoulos, Priyanka Suresh, Bailey Herms, Justin Grayston, Benjamin Maynard, Nicholas Dietrich, Robbe Wulgaert, Jonathan Prag, Alex Mullen, and Shakir Mohamed. 2025. Contextualizing ancient texts with generative neural networks. *Nature*, 645(8079):141–147.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, et al. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

William H. Baxter and Laurent Sagart. 2014. *Old Chinese: A New Reconstruction*. Oxford University Press.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

CASS. 2007. *Yin Zhou Jin Wen Ji Cheng (Complete Collection of Yin and Zhou Bronze Inscriptions (CCYZBI))*. Zhonghua Shuju (Zhonghua Book Company), Beijing.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Danlu Chen, Jiahe Tian, Yufei Weng, Taylor Berg-Kirkpatrick, and Jacobo Myerston. 2024. Classification of paleographic artifacts at scale: Mitigating confounds and distribution shift in cuneiform tablet dating. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 30–41, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Mengjia Chen. 2004. *Xi Zhou Tongqi Duandai (Chronology of Western Chou Bronze Vessels)*. Zhonghua Shuju (Zhonghua Book Company), Beijing.

Ya-Min Chou and Chu-Ren Huang. 2005. 異體字語境關係的分析與建立 (a framework for the contextual analysis of Chinese characters variants) [in Chinese]. In *Proceedings of the 17th Conference on Computational Linguistics and Speech Processing*, pages 273–291, Tainan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinqin Du. 2020. Shang dai jin wen tong yong zi zheng li yu yan jiu (arrangement and research of common characters in bronze in shang dynasty). Master's thesis, Southwest University, Chongqing.

Sijia Ge. 2022. Integration of named entity recognition and sentence segmentation on Ancient Chinese based on siku-BERT. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 167–173, Taipei, Taiwan. Association for Computational Linguistics.

Moruo Guo. 1999. *Liang Zhou Jin Wen Ci Da Xi Tu Lu Kao Shi*. Shanghai Shudian (Shanghai Bookstore Publishing House), Shanghai.

Rui Guo. 2021. A research on an intelligent recognition tool for bronze inscriptions of the shang and zhou dynasties. *Journal of Chinese Writing Systems*, 4(4):271–279. Original work published 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Wenjie Hua and Shenghan Xu. 2025. When less is more: Logits-constrained framework with RoBERTa for Ancient Chinese NER. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 192–196, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Hai Huang. 2022. *Hu Ding Tong Kao*. Gezhi Chubanshe (Truth & Wisdom Press), Shanghai.

Rongquan Jin. 2014. Zhou dai fan guo qing tong qi ji qi li shi di li lun kao(on zhou period fan state bronzes, history and geography). *Huaxia Kaogu (Huaxia Archaeology)*, 2:62.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chuntao Li. 2024. Ren gong zhi neng yu jin wen yan jiu zhan wang. *Chinese Social Sciences Today*. `https://www.cssn.cn/skgz/bwyc/202408/t20240809_5769948.shtml`.

Jinhao Li, Zijian Chen, Runze Jiang, Tingzhu Chen, Changbo Wang, and Guangtao Zhai. 2025. Mitigating long-tail distribution in oracle bone inscriptions: Dataset, model, and benchmark.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhiji Liu. 2009. Jian shuo gu wen zi yi ti zi de fa zhan yan bian (on the development and evolution of ancient varient forms of chinese characters). In *Zhongguo Wenzi Yanjiu (The Study of Chinese Characters)*, volume 12, pages 36–46. Daxiang Chubanshe (Elephant Press), Zhengzhou.

Tingting Luo. 2013. Dong zhou jin wen tong jia zi yan jiu (interchangeable characters in bronze inscriptions of eastern zhou dynasty). Master's thesis, Yunnan University, Kunming.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Chengyuan Ma. 1986. *Shang Zhou Qing Tong Qi Ming Wen Xuan*. Wenwu Chubanshe (Cultural Relics Press), Beijing.

Chengyuan Ma. 2003. *Zhong Guo Qing Tong Qi*. Shanghai Guji Chubanshe (Shanghai Classics Publishing House), Shanghai.

Dimitrios Meletis. 2019. The grapheme as a universal basic unit of writing. *Writing Systems Research*, 11(1):26–49.

Dimitrios Meletis. 2020. Types of allography. *Open Linguistics*, 6:249–266.

Kamline Miloud, Moulay Lakhdar Abdelmounaim, Beladgham Mohammed, and Bendjillali Ridha Ilyas. 2024. Restoration of ancient arabic manuscripts: A deep learning approach. *Studies in Engineering and Exact Sciences*, 5(2):1–22.

Bofeng Mo, Weiqi Qiu, and Zecheng Xie. 2021. Ren Gong Zhi Neng Mo Ni Ci Li Gui Na De Chu Bu Ce Shi (preliminary test of artificial intelligence simulation speech case induction). *Han Yuyan Wenxue Yanjiu (Chinese Language and Literature Research)*, 12(3):128–135.

Hoang Nguyen, Chenwei Zhang, Congying Xia, and Philip S Yu. 2020. Dynamic semantic matching and aggregation network for few-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1209–1218.

Hoang Nguyen, Chenwei Zhang, Tao Zhang, Eugene Rohrbaugh, and Philip Yu. 2023. Enhancing cross-lingual transfer via phonemic transcription integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9163–9175, Toronto, Canada. Association for Computational Linguistics.

Hoang H Nguyen, Khyati Mahajan, Vikas Yadav, Julian Salazar, Philip S. Yu, Masoud Hashemi, and Rishabh Maheshwary. 2025. Prompting with phonemes: Enhancing LLMs' multilinguality for non-Latin script languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11975–11994, Albuquerque, New Mexico. Association for Computational Linguistics.

Song-Liang Pan, Kunchi Li, Da-Han Wang, Xu-Yao Zhang, Jiantao Liu, and Shunzhi Zhu. 2026. Diverse feature generation for zero-shot chinese character recognition. *Expert Systems with Applications*, 297:129442.

Ruihua Qi. 2023. Xi zhou jin wen tong jia guan xi zheng li yu yan jiu (the collation and study of tongjia in the bronze inscriptions of the western zhou). Master's thesis, Jilin University, Changchun.

Xigui Qiu. 2012. *Qiu Xi Gui Xue Shu Wen Ji (Collected Works of Qiu Xigui)*. Fudan University Press, Shanghai.

Xigui Qiu. 2013. *Wen Zi Xue Gai Yao (The Essentials of Grammatology)*. Shangwu Yinshuguan (The Commercial Press), Beijing.

Shizuka Shirakawa. 1962. *Kinbun tsūshaku, Vols. 1-10; Hakutsuru Bijutsukan shi, Vols. 1-56*. Hakutsuru Bijutsukan, Kobe.

Wenying Su. 2016. *Xi Zhou Jin Wen Yi Ti Zi Yan Jiu (Research on Variant Characters of Inscriptions on Bronze Objects during the Dynasty of Western Zhou)*. Phd dissertation, Southwest University, Chongqing.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Lan Tang. 2016. *Xi Zhou Qing Tong Qi Ming Wen Fen Dai Shi Zheng*. Shanghai Guji Chubanshe, Shanghai.

Zuoyu Tian and Sandra Kübler. 2021. Period classification in Chinese historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 168–177, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Dongbo Wang, Chang Liu, Zihe Zhu, Jangfeng Liu, Haotian Hu, Si Shen, and Bin Li. 2021. Construction and application of pre-trained models of siku quanshu in orientation to digital humanities. *Library Tribune*.

Shuai Wang. 2015. *Xi Zhou Jin Wen Zi Xing Shu Ti Yan Bian Yan Jiu Yu Tong Qi Duan Dai (A Study of Figure of Inscriptions on Ancient Bronzes and Dating of Bronze Vessels of the Western Zhou Period)*. Phd dissertation, Shaanxi Normal University, Xi'an.

Zhaoji Wang, Shirui Zhang, Xuetao Zhang, and Renfen Hu. 2023. 古通假字源的构建及用研究(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.

Zhen Wang, Yujun Li, and Honglei Li. 2025. Chinese inscription restoration based on artificial intelligent models. *npj Heritage Science*, 13(1):326.

Zhenfeng Wu. 2012. *Shang Zhou Qing Tong Qi Ming Wen Ji Tu Xiang Ji Cheng*. Shanghai Guji Chubanshe (Shanghai Classics Publishing House), Shanghai.

Zhenyu Wu. 2023. *Liang Zhou Jin Wen Yu Fa Yan Jiu*. Shangwu Yinshuguan (The Commercial Press), Beijing.

Mingwen Xie. 2014. Tan tan jin wen zhong song ren suo wei "zhi" de zi ming. https://www.fdgwz.org.cn/Web/Show/2406#_ednref5. Center for Research on Chinese Excavated Classics and Paleography at Fudan University, accessed 2014-12-25.

Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. 2023. Chinese text recognition with a pre-trained clip-like model through image-ids aligning.

Lingbin Zeng. 2011. Hubei suizhou ye jia shan xi zhou mu di fa jue jian bao. *Wen Wu (Cultural Relics)*, 11:31.

Linhan Zhang, Qian Chen, Wen Wang, Chong Deng, Xin Cao, Kongzhang Hao, Yuxin Jiang, and Wei Wang. 2023. Weighted sampling for masked language modeling. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

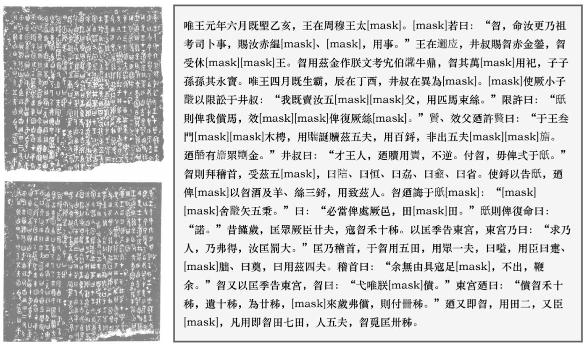
Ruo Qing Zhao, Hui Qin Wang, Ke Wang, et al. 2020. Recognition of bronze inscriptions image based on mixed features of histogram of oriented gradient and gray level co-occurrence matrix. *Ji Guang Yu Guang Dian Xue Jin Zhan (Laser & Optoelectronics Progress)*, 57(12):98–104.

Mask Position	Gold	Pred@1	Top5
01	室	廟	廟 室 宮 寢 廷
02	王	王	王 公 君 伯 尹
03	帯	帯	帯 純 衡 衣 韍
05	命	於	於 于 揚 無 多
06	于	于	于 揚 穆 於 侑
07	年	年	年 人 世 壽 歲

Table 5: Top-1 and Top-5 predictions versus gold characters (excerpt of the first six damaged positions in the *Hu Ding* inscription).

A Case Study: *Hu Ding* Restoration

To further illustrate our approach, we applied the model to the mid–Western Zhou *Hu Ding* bronze vessel (CCYZBI.02838A/B), a well-known inscriptional source with multiple damaged positions. To avoid leakage, we excluded all *Hu Ding* entries from BIRD before training. The task is thus strictly out-of-sample.



唯王元年六月既朢乙亥，王在周穆王太[mask]。[mask]若曰：“曶，命汝更乃祖考司卜事，賜汝赤緼[mask]、[mask]，用事。”王在邐应，井叔賜曶赤金鋚，曶受休[mask][mask]王。曶用茲金作朕文考宄伯䵼牛鼎，曶其萬[mask]用祀，子子孫孫其永寶。唯王四月既生霸，辰在丁酉，井叔在異為[mask]。[mask]使厥小子𩣑以限訟于井叔：“我既賣汝五[mask][mask]父，用匹馬束絲。”限許曰：“䛅則俾我償馬，效[mask][mask]俾復厥絲[mask]。”賢、效父迺許賫曰：“于王叄門[mask][mask]木榜，用𧶠誕贖茲五夫，用百鋝，非出五夫[mask][mask]𤸫。迺𩢺有𤸫罘𩰫金。”井叔曰：“才王人，迺贖用賁，不逆。付曶，毋俾式于䛅。”曶則拜稽首，受茲五[mask]，曰陪、曰恒、曰劦、曰𨪰、曰省。使鋝以告䛅，迺俾[mask]以曶酒及羊、絲三鋝，用致茲人。曶迺誨于䛅[mask]：“[mask][mask]舍𩣑矢五秉。”曰：“必當俾處厥邑，田[mask]田。”䛅則俾復命曰：“諾。”昔饉歲，匡眾厥臣廿夫，寇曶禾十秭。以匡季告東宮，東宮乃曰：“求乃人，乃弗得，汝匡罰大。”匡乃稽首，于曶用五田，用眾一夫，曰嗌，用臣曰疐、[mask]朏、曰奠，曰用茲四夫。稽首曰：“余無由具寇足[mask]，不出，鞭余。”曶又以匡季告東宮，曶曰：“弋唯朕[mask]償。”東宮迺曰：“償曶禾十秭，遺十秭，為廿秭，[mask]來歲弗償，則付卌秭。”迺又即曶，用田二，又臣[mask]，凡用即曶田七田，人五夫，曶覓匡卅秭。

Figure 4: Left: Rubbing of the *Hu Ding* inscriptions (CCYZBI.02838A, 02838B) (CASS, 2007), image courtesy of *AS DABII*. Right: Transcription from (Huang, 2022), used as the input with damaged positions masked.

We employed the SikuRoBERTa (GN) model with two decoding strategies: parallel mask filling and greedy iterative decoding (Lazar et al., 2021). Table 5 compares predicted tokens with expert gold restorations.

On 22 expert restorations (Huang, 2022), the model achieved Exact@1: 50.00% (11/22), Exact@5: 59.09% (13/22), and Exact@10: 68.18% (15/22) under the parallel prediction setting. Greedy decoding yielded comparable coverage, though with a lower accuracy. In addition to reproducing expert restorations, the system generated plausible candidates for characters that remain undeciphered (Table 6), providing a potential reference for paleographic analysis.

Mask Position	Top10 Predictions
04	鑾 旂 舄 筆 U+3AC3 金 矢 黃 弓 璋
08	介 伯 市 限 客 期 制 政 宰 人
15	之 外 一 若 內 賜 邑 大 下 又
16	賜 折 喬 杜 乘 造 擇 柞 之 于
17	則 許 弗 不 人 亦 也 而 帛 乃
18	則 曰 不 弗 許 告 厥 多 有 用
28	其 厥 若 越 乃 我 以 汝 如 余

Table 6: Model completions for undeciphered positions in *Hu Ding* (Top-10 shown)

B DAPT Composition

The DAPT (Pre-Qin) corpus consists of 40 transmitted and excavated texts, compiled and normalized from open sources. Following the classification of the *Chinese Text Project*, Table 7 presents a subset categorized. These texts provide broad coverage of syntactic and lexical patterns closely aligned with inscriptional Chinese.

C Training Objective

Our model training combines three terms:

Masked language modeling (MLM). The base loss is the standard MLM objective:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i\in\mathcal{M}} \log P_\theta(y_i \mid X_i),$$

where $\mathcal{M}$ is the set of masked positions, y_i the gold glyph, and X_i the masked context.

Glyph-biased masking. To emphasize historically informative glyphs, candidate positions are sampled with bias:

$$p(i) = \frac{w_i}{\sum_{j\in\mathcal{C}} w_j}, \quad w_i = \begin{cases} \lambda & i \in \mathcal{G}, \\ 1 & \text{otherwise}, \end{cases}$$

where $\mathcal{C}$ is the set of candidate tokens, $\mathcal{G}$ the set of glyph tokens in GN clusters, and $\lambda > 1$ controls the bias strength.

Glyph-net (GN) regularization. To encourage consistency across allographs, for each cluster G with token set $\mathcal{T}(G)$ we define:

$$\mathcal{L}_{\text{GN}} = -\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \frac{1}{|\mathcal{T}(G_i)|}\sum_{t\in\mathcal{T}(G_i)} \log P_\theta(t \mid X_i).$$

Final objective. The training loss interpolates between MLM and GN terms:

$$\mathcal{L} = (1-\alpha)\,\mathcal{L}_{\text{MLM}} + \alpha\,\mathcal{L}_{\text{GN}},$$

with α gradually scheduled during training.

Category	Titles
Confucianism	*The Analects*, *Mengzi*, *Liji*, *Xiao Jing*, *Xunzi*
Mohism	*Mozi*
Daoism	*Dao De Jing*, *Zhuangzi*, *Liezi*, *He Guan Zi*
Legalism	*Hanfeizi*, *Shang Jun Shu*, *Shenzi*, *Jian Zhu Ke Shu*, *Guanzi*
School of Names	*Gongsunlongzi*
School of the Military	*The Art of War*, *Wu Zi*, *Liu Tao*, *Si Ma Fa*, *Wei Liao Zi*
Miscellaneous Schools	*Lü Shi Chun Qiu*, *Gui Gu Zi*
Histories	*Chun Qiu Zuo Zhuan*, *Lost Book of Zhou*, *Guo Yu*, *Yanzi Chun Qiu*, *Zhan Guo Ce*, *Zhushu Jinian*, *Mutianzi Zhuan*
Ancient Classics	*Book of Poetry*, *Shang Shu*, *Book of Changes*, *Rites of Zhou*, *Chu Ci*, *Yili*, *Shan Hai Jing*
Medicine	*Huangdi Neijing*
Excavated	*Guodian*, *Mawangdui*

Table 7: Subset of Pre-Qin texts in the DAPT corpus. The Mawangdui Silk Texts, though excavated from a Western Han tomb (archaeological date), reflect Pre-Qin synchronic register (compositional date) and are thus included.

D Training Setup

We evaluate four adaptation schedules: (i) no adaptation, (ii) domain-adaptive pretraining (DAPT), (iii) task-adaptive pretraining (TAPT), and (iv) a two-stage DAPT→TAPT pipeline. Each schedule is optionally combined with Glyph Net alignment or glyph-biased sampling. All baselines are extended with UNK placeholders for unseen characters. In DAPT, the bottom six layers are frozen and trained for ten epochs on a large Pre-Qin corpus; TAPT then unfreezes all layers and adapts to inscriptional data. The two stages are interpolated with a weighting parameter λ balancing DAPT and TAPT losses.

E Hyper-parameters

We found the best hyperparameters for each model during the search via Weights & Biases (Biewald, 2020), as detailed in Table 8.

Hyper-parameter	mBERT	XLM-Base	XLM-Large	SikuRoBERTa
Learning Rate	0.00005	0.00005	0.00005	0.00012
Epochs	60	40	40	40
Batch Size	32	32	32	32
Attention Dropout	0.1	0.1	0.1	0.1
Hidden Dropout	0.1	0.1	0.1	0.1
Stride	12	10	12	10
mlm_prob	0.2	0.2	0.2	0.2
Weight Decay	0.01	0.01	0.01	0.01

Table 8: Best hyperparameters found during WandB hyperparameter search for mBERT, XLM-Base, XLM-Large, and SikuRoBERTa.

F Undeciphered Characters

Figure 5 shows glyphs from bronze inscriptions that remain undeciphered by paleographers. The complete collection of undeciphered forms can be found in our GitHub repository. We encode these glyphs with symbolic placeholders, assigning a distinct identifier to each form. Visually identical forms are mapped to the same identifier.

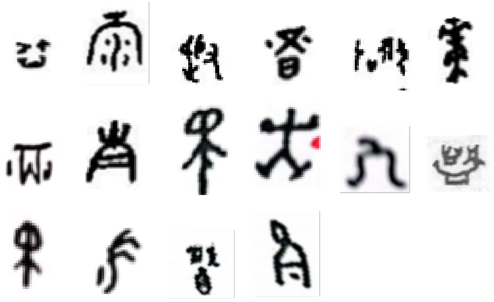


Figure 5: Examples of undeciphered glyphs represented by UNK placeholders in BIRD.

G Paleographical References

We draw on recent paleographical and historical studies of bronze inscriptions to update character forms and chronological assignments in our corpus. For example, an inscription previously dated to the Early Spring and Autumn period (CCYZBI.02737 (CASS, 2007)) has been reassigned to the Middle Spring and Autumn period in (Wu, 2012); similarly, another item formerly placed in the Middle Western Zhou (CCYZBI.02737 (CASS, 2007)) has been revised to the Middle Spring and Autumn period in (Jin, 2014). For further details, please refer to our GitHub repository, which will continue to be updated in the future.

H Ablation

We conduct ablation studies across four backbones (SikuRoBERTa, mBERT, XLM-Base, XLM-Large) to disentangle the effects of domain- and task-adaptive pretraining, allograph-aware supervision, and glyph-biased sampling. Restoration accuracy is summarized in Table 9, dating performance is reported in Table 10, and representation cohesion and separation are analyzed in Table 11.

Model	Scenario	E@1 ↑	E@5 ↑	E@10 ↑	F@1 ↑	F@5 ↑	F@10 ↑
SikuRoBERTa	Baseline	0.236	0.377	0.440	0.244	0.395	0.458
	DAPT_only	0.260	0.423	0.494	0.253	0.432	0.512
	TAPT_Bias	0.483	0.626	0.676	0.544	0.678	0.731
	TAPT_GN	**0.495**	**0.652**	**0.702**	0.543	0.681	**0.731**
	TAPT_GN_Bias	0.492	0.638	0.686	**0.554**	**0.688**	0.729
	TAPT_from_DAPT	0.485	0.636	0.685	0.535	0.681	0.729
	TAPT_only	0.488	0.639	0.684	0.539	0.681	0.723
mBERT	Baseline	0.112	0.224	0.282	0.093	0.205	0.267
	DAPT_only	0.148	0.283	0.353	0.139	0.278	0.355
	TAPT_Bias	0.427	0.572	0.622	0.464	0.617	0.665
	TAPT_GN	**0.436**	**0.586**	**0.637**	**0.469**	0.613	0.659
	TAPT_GN_Bias	0.424	0.583	0.635	0.466	**0.618**	**0.665**
	TAPT_from_DAPT	0.431	0.574	0.623	0.464	0.607	0.657
	TAPT_only	0.427	0.570	0.613	0.465	0.606	0.648
XLM-Base	Baseline	0.122	0.195	0.234	0.112	0.187	0.228
	DAPT_only	0.161	0.279	0.337	0.151	0.270	0.332
	TAPT_Bias	0.432	0.572	0.622	**0.454**	0.598	0.644
	TAPT_GN	**0.435**	**0.584**	**0.629**	0.443	0.595	0.640
	TAPT_GN_Bias	0.429	0.583	0.626	**0.454**	**0.608**	**0.651**
	TAPT_from_DAPT	0.434	0.583	0.629	0.447	0.595	0.639
	TAPT_only	0.424	0.557	0.602	0.434	0.568	0.614
XLM-Large	Baseline	0.140	0.225	0.265	0.132	0.208	0.257
	DAPT_only	0.178	0.321	0.382	0.166	0.312	0.384
	TAPT_Bias	0.453	0.595	0.640	**0.479**	**0.615**	0.656
	TAPT_GN	**0.456**	**0.609**	**0.649**	0.472	0.612	0.654
	TAPT_GN_Bias	0.454	0.598	0.645	0.476	0.609	**0.657**
	TAPT_from_DAPT	0.456	0.600	0.648	0.471	0.604	0.649
	TAPT_only	0.435	0.577	0.621	0.442	0.584	0.622

Table 9: Restoration results under different adaptation schedules across four pretrained models. E@k denotes Exact@k and F@k denotes Family@k. All values are reported as proportions between 0 and 1. Best results per column are bolded.

Model	Scenario	Acc_Dyn ↑	F1_Dyn ↑	Acc_Hier_Dyn ↑	F1_Hier_Dyn ↑	Acc_Hier_Per ↑	F1_Hier_Per ↑
SikuRoBERTa	DAPT_only	0.833	0.728	0.836	**0.552**	**0.684**	0.630
	TAPT_only	0.846	0.727	**0.849**	0.570	0.651	0.605
	TAPT_from_DAPT	0.840	0.698	0.836	0.544	0.671	0.627
	TAPT_GN	0.840	0.698	0.842	0.542	**0.684**	**0.638**
	TAPT_GN_Bias	0.852	0.767	0.849	0.539	0.678	0.635
	TAPT_Bias	**0.864**	**0.778**	0.842	0.543	0.671	0.629
mBERT	Baseline	0.809	0.672	0.822	0.515	0.638	0.583
	TAPT_only	**0.846**	**0.762**	**0.836**	0.540	**0.664**	**0.616**
	TAPT_from_DAPT	**0.846**	0.752	0.822	0.534	0.658	**0.616**
	TAPT_GN	**0.846**	0.745	0.822	0.534	0.618	0.575
	TAPT_GN_Bias	0.815	0.642	0.829	**0.547**	0.651	0.601
	TAPT_Bias	**0.846**	0.748	0.822	0.531	0.638	0.586
XLM-Base	Baseline	0.673	0.277	0.763	0.379	0.592	0.520
	DAPT_only	0.778	0.483	0.796	0.493	0.625	0.567
	TAPT_only	0.778	0.476	0.789	0.498	0.612	0.566
	TAPT_from_DAPT	0.784	0.486	**0.809**	0.484	0.618	**0.576**
	TAPT_GN	0.765	0.429	0.803	0.433	**0.632**	0.569
	TAPT_GN_Bias	0.784	0.481	**0.809**	0.502	0.605	0.556
	TAPT_Bias	**0.790**	**0.503**	**0.809**	**0.513**	0.625	0.573
XLM-Large	Baseline	0.747	0.444	0.803	0.436	0.618	0.542
	DAPT_only	0.772	0.566	0.822	0.540	0.612	0.580
	TAPT_only	0.815	0.667	**0.849**	0.572	**0.678**	**0.657**
	TAPT_from_DAPT	0.809	0.655	**0.849**	**0.581**	0.658	0.630
	TAPT_GN	0.821	0.705	0.809	0.512	0.572	0.534
	TAPT_GN_Bias	**0.840**	0.701	0.836	0.526	0.612	0.561
	TAPT_Bias	**0.840**	**0.746**	0.816	0.531	0.651	0.630

Table 10: Classification results for dynasty- and period-level dating under different adaptation schedules. Acc = accuracy, F1 = macro-F1. **Dyn** = dynasty-level classification (single-task); **Hier_Dyn** = dynasty-level accuracy/F1 in the hierarchical model; **Hier_Per** = period-level accuracy/F1 in the hierarchical model, where period prediction is conditioned on the predicted dynasty. Best results per column are bolded.

Model	Scenario	IntraCos Avg (↑)	Nearest-InterCos Avg (↓)
SikuRoBERTa	Baseline	0.494	0.252
	DAPT_only	0.488	0.266
	TAPT_Bias	0.503	0.292
	TAPT_GN	**0.515**	0.291
	TAPT_GN_Bias	0.514	0.290
	TAPT_from_DAPT	0.504	0.295
	TAPT_only	0.486	**0.255**
mBERT	Baseline	0.496	0.218
	DAPT_only	0.470	**0.197**
	TAPT_Bias	0.483	0.215
	TAPT_GN	0.492	0.219
	TAPT_GN_Bias	**0.493**	0.220
	TAPT_from_DAPT	0.482	0.215
	TAPT_only	0.471	0.199
XLM-Base	Baseline	0.516	0.309
	DAPT_only	0.522	0.317
	TAPT_Bias	0.551	0.349
	TAPT_GN	0.553	0.349
	TAPT_GN_Bias	0.553	0.349
	TAPT_from_DAPT	**0.557**	0.356
	TAPT_only	0.519	**0.313**
XLM-Large	Baseline	0.530	0.342
	DAPT_only	0.532	0.345
	TAPT_Bias	0.553	0.366
	TAPT_GN	0.554	0.365
	TAPT_GN_Bias	**0.555**	0.367
	TAPT_from_DAPT	0.554	0.366
	TAPT_only	0.532	**0.344**

Table 11: Representation analysis of allograph clusters. *IntraCos Avg* (↑) measures within-cluster cohesion by averaging cosine similarity between tokens and their cluster centroids. *Nearest-InterCos Avg* (↓) measures between-cluster separation by averaging the cosine similarity of each cluster to its nearest neighbor. Together, indicate how well the embedding space encodes palaeographic grapheme-allograph structure. Best results per column are bolded.