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
Wuhan University
gegangyan@163.com

Abstract

Bronze inscriptions from early China are often fragmentary, with missing or undeciphered characters and uncertain chronological assignments. To address this, we propose **BIRD** (Bronze Inscription Restoration and Dating), a dataset and framework that leverages pretrained language models (PLMs) tailored to the unique demands of ancient texts. By integrating domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT) techniques, along with a glyph net resource that links graphemes and allographs, our approach overcomes key challenges in low-resource settings and the prevalence of allography. Our results show marked improvements in both restoration and dating accuracy.

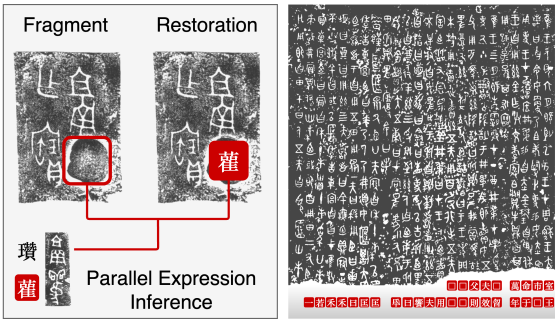

Figure 1: Left: A simplified paleographer's workflow for restoring a damaged bronze inscription: identifying the damaged fragment, inferring from parallel expressions, and proposing a restoration (Zeng, 2011; Wu, 2012; Xie, 2014). Right: A damaged bronze inscription fragment (CCYZBI.02838A) (CASS, 2007) with expert annotations (Huang, 2022).

1 Introduction

Bronze inscriptions from the Chinese Bronze Age (c. 21st–3rd century 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 graphic forms and contextual inference, a process that is both labor-intensive and difficult to scale. Neural models, particularly pretrained 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.

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

1. **Low-resource setting.** Although nearly 20,000 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 hinder-

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 (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*).

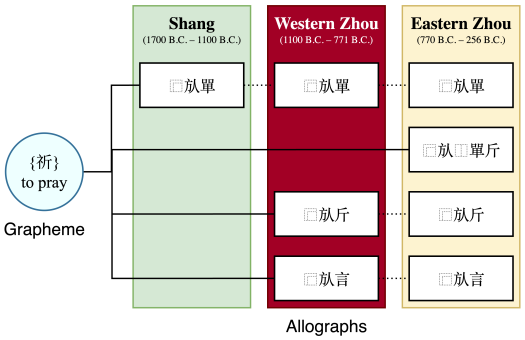

Figure 2: Concrete glyph family of *Qi* ('to pray') from the Shang to the Eastern Zhou. To illustrate the correlation between glyphs and their components, Ideographic Description Sequences (IDS) are used.

ing generalization in data-hungry Transformers. Figure 2 shows a representative family of allographs that share the same grapheme.

Nonetheless, bronze inscriptions also share their linguistic environment with transmitted and excavated Pre-Qin (c. 21st–3rd century BCE) texts (Li, 2024), which can serve as auxiliary corpora for domain-adaptive pretraining (Gururangan et al., 2020). Moreover, allographic variation is not mere noise: in downstream tasks such as chronological dating, these glyph distinctions provide evidence (Wang, 2015; Su, 2016). Effective modeling must therefore balance normalization for learning with preservation of 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 a glyph net (GN), a resource that pairs and clusters graphemes and their allographs into sets;
- A variant-aware masked language modeling framework for character restoration and downstream dating.

2 Related Work

Scholarly work on bronze texts has a long history. Paleographically, CASS (2007) provides a comprehensive compilation, while Wu (2012), Ma (1986), and Shirakawa (1962) offer translations and philological interpretations. Chronological dating has also been extensively studied, with theories and 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, and lexica. 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 on variant families across periods (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.

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.

3.2 BIRD (TAPT)

For bronze inscriptions, existing resources such as *Jihewang*, *AS DABII*, and *AS Lexicon* function primarily as retrieval platforms rather than structured datasets. We therefore release a dataset that addresses these gaps.

We present the first bronze inscription dataset designed for NLP applications such as restoration and dating. BIRD contains 41k tokens and is accompanied by 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 that graphemes and their allographs are mutually substitutable (Qiu, 2012).

Inscriptions are also labeled with dynasty and finer-grained period, which allows supervised experiments on dating. Table 1 compares BIRD with previous digitization efforts.

The corpus itself is prepared through four steps:

1. **Encoding.** All inscriptions are converted into machine-readable text. As shown in Table 2, characters are categorized as (i) identifiable, (ii) damaged and unreadable (marked as □), or (iii) visually legible but undeciphered (encoded as `[UNK-xxxxx-x]`). See Appendix F.
2. **Filtering.** Extremely short inscriptions ($\leq$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 and risk leakage, so we retain only one representative instance.
4. **Correction.** Clerical transcriptions (*liding*) and chronological assignments are updated in line with recent philological research, with dynasty and period labels attached to each inscription. Refer to Appendix G for more information.

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 contemporaneous a Pre-Qin corpus, with shallow layers frozen to stabilize training; (2) a glyph net (GN) constructed from grapheme–allograph pairs, where transitive closure yields glyph families and new glyphs are aligned to family centroids; (3) GN-aware masking and substitution, which bias token sampling toward glyph tokens and encourage in-family replacements, thereby promoting family-level learning. These modifications preserve the effectiveness of MLM training while fully exploiting glyph information to benefit downstream tasks. 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 BERT and RoBERTa have demonstrated strong transfer performance in low-resource and cross-lingual settings (Lazar et al., 2021; Chau et al., 2020), while

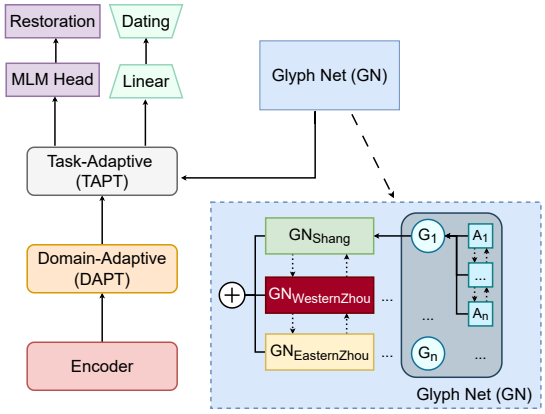

Figure 3: Our pipeline enhances masked language modeling for bronze inscriptions by combining domain-adaptive pretraining (DAPT), task-adaptive pretraining (TAPT), and a Glyph Net module (as illustrated in the lower-right component, each grapheme $G_{1..n}$ is linked to its allographs $A_{1..n}$) that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

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 (median length four characters in BIRD), so standard BERT masking often erases all context. We instead use stride-based masking (s), ensuring that sequences of length $\leq s$ lose at most one token. The stride is tuned per backbone via Bayesian 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	**48.50**	**63.59**	**68.47**	53.52	**68.82**	**72.86**
mBERT	110M	42.42	58.30	63.52	46.62	61.77	66.46
XLM-Base	278M	42.91	58.25	62.63	45.38	60.78	65.10
XLM-Large	550M	45.35	59.78	64.47	47.58	60.85	65.65

Table 3: Restoration results on zero-shot held-out glyph forms. E@K = Exact@K; F@K = Family@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. Every s-th non-boundary character is masked, and performance is measured by: *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 reported results are zero-shot on held-out glyph forms excluded from training. This setup simulates the real conditions of restoration.

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 48.50 Exact@1 and 72.86 Family@10, outperforming BiLSTM by +9.5 p.p. (Exact@1) and +10.4 p.p. (Family@10). BiLSTM only leads on Family@1 (57.41). Multilingual PLMs lag SikuRoBERTa by 3–6 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. All backbones trained on bronze-domain text substantially outperform the uniform random baseline. SIKUROBERTA achieves the strongest results overall (dynasty accuracy 85.19; macro-F1 76.69) and the highest hierarchical period F1 (63.55). MBERT lags by 2–12 absolute points across metrics, while larger multilingual encoders (XLM-BASE/LARGE) fail to match domain-adapted performance; for XLM-BASE, learning fine-grained

Model	Params	Dynasty		Hier-Dyn		Hier-Per	
		Acc	F1	Acc	F1	Acc	F1
Random	–	25.00	–	25.00	–	8.33	–
SVM	0.08M	75.31	49.44	76.32	42.67	58.55	49.43
SikuRoBERTa	109M	**85.19**	**76.69**	**84.87**	53.95	**67.76**	**63.55**
mBERT	110M	81.48	64.25	82.89	**54.67**	65.13	60.06
XLM-Base	278M	78.40	48.13	80.92	50.24	60.53	55.57
XLM-Large	550M	83.95	70.12	83.55	52.58	61.18	56.09

Table 4: Dating performance across baselines and pretrained models. **Dynasty** = four-way classification; **Hier-Dyn / Hier-Per** = hierarchical evaluation at dynasty and period levels. Random baselines: 25.00 (dynasty), 8.33 (period). All scores are percentages.

period distinctions modestly improves dynasty prediction. Overall, period classification remains more challenging than dynasty classification.

6.4 Analysis

For restoration, through correct predictions, we find that the model has learned relatively stable conditional distributions. Restoration is especially strong in formulaic segments of bronze inscriptions, where contextual patterns are fixed. Nouns denoting vessels, temporal adverbs, and modal particles, elements with syntactically fixed functions (Wu, 2023), are likewise restored with high accuracy. Overall, due to the stereotyped and narrative-regular nature of bronze texts (Ma, 2003), characters embedded in stable co-occurrence contexts are easier to restore.

On the error side, the model tends to default to high-frequency template positions when uncertain. Confusions also arise among official titles and kinship nouns, within prepositional groups introducing causes or objects, and across semantically related action verbs. Numerals are mutually confusable, which reflects the lack of fine-grained discriminative signals. Nevertheless, even in mispredictions, the predicted top candidates often fall into the correct syntactic category, indicating grammatical awareness.

For dating, common errors across models are that they tend to overfit on official titles while neglecting chronological distinctions. In terms of the overall error distribution, the Spring and Autumn and Warring States periods are less stable, which corresponds to the textual feature of inscriptions from these eras being relatively free in style (Ma, 2003). In addition, formulaic expressions are shared across periods, which blurs the boundaries between dynasties. Under conditions where the majority class dominates the data, misclassifications are more likely to be pulled toward Western Zhou. Nevertheless, it is essential to note that severe misplacements across distant periods are relatively rare, and the accuracy for the majority class remains very high, which reveals that the models have indeed learned effective chronological signals from the data.

7 Conclusion

We present BIRD (Bronze Inscription Restoration and Dating), a dataset and an approach to restoring and dating by leveraging masked language models (MLMs) adapted to the unique challenges posed by these ancient texts. The results demonstrate that our proposed framework, particularly on the SikuRoBERTa backbone, performs well and provides substantial improvements in restoration and dating tasks. By integrating domain-adaptive pretraining (DAPT), task-adaptive pretraining (TAPT), and glyph net-aware training, we achieve high accuracy in both tasks. This work sets a foundation for future NLP applications in 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 network currently relies merely on group-level inductive bias. Its generalization could be strengthened by incorporating stricter palaeographic constraints (Chou and Huang, 2005) and expanding knowledge bases 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 phonetic 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. Many inscriptions are partially legible, with subcomponents visible even when the full graph is damaged. A token-level MLM cannot leverage such partial signals. Structure-aware encodings such as Ideographic Description Sequences (IDS), 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 archaeological and multimodal signals that are central to traditional chronology. Vessel shape, decorative motifs, and casting techniques provide independent chronological evidence (Chen, 2004), yet remain unexploited. Integrating textual modeling with such modalities would bring the system closer to expert palaeographic practice.

Finally, it is important to clarify the positioning of this work. BIRD provides the first fully encoded, NLP-ready dataset and baseline framework for bronze inscription restoration and dating. As such, its role is primarily to supply standardized resources and computational baselines for future research, rather than to replace traditional philological methods. Current Language Model-based predictions should be regarded only as auxiliary hypotheses, offering preliminary guidance for palaeographers. Ultimately, expert interpretation grounded in palaeographic and archaeological context remains 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, 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.

Ruo Qing Zhao, Hui Qin Wang, Ke Wang, et al. 2020. Ji yu fang xiang ti du zhi fang tu he hui du gong sheng ju zhen hun he te ding de jin wen tu xiang shi bie (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
1	室	室	室廟宮寢廷
2	王	王	王君公伯尹
3	市 (or 芾)	芾	芾衣衡純戈
4	Unde	鑾	鑾旂馬舄車
5	命	亡	亡無有多毋
6	于	宜	宜告御命掖
7	年	年	年人世壽
8	Unde	伯	伯室守圃大

Table 5: Greedy (Top-1) predictions versus gold characters (excerpt of the first 8 positions). **Unde** denotes undeciphered characters. Parenthetical alternatives **or** reflect variant readings attested in other expert transcriptions (CASS, 2007).

A Case Study: *Hu Ding* Restoration

We applied our model to restore a partially damaged inscription from the *Hu Ding* bronze vessel, dating back to the mid-Western Zhou period. The transcription of the inscription is shown in Figure 4.

唯王元年，六月既望乙亥，王在周穆王太[mask]。[mask]若曰：曾，命汝更乃祖考司卜事，賜汝赤舄[mask]、[mask]，用事。”王在邁应，井叔賜曾赤金、鑾，曾受休[mask][mask]于王。曾用茲金作联文考宄伯噂牛鼎，曾其萬[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]使厥小子陪以限訟于井叔：“我既賓汝五[mask][mask]父，用匹馬束絲。”限許曰：氐則俾我僦馬，效[mask][mask]俾復厥絲[mask]，賢、效父遘許晉曰：“于王參門[mask][mask]木楞，用賠誕曠茲五夫，用百鋝，非出五夫[mask][mask]旂。遘卲有旂累鋝金。”井叔曰：“才王人，遘曠用罰，不逆。付曾，毋俾式于氐。”曾則拜稽首，受茲五[mask]，曰笏、曰恒、曰荔、曰愈、曰省，使鋝以告氐，遘俾[mask]以曾酒及羊、絲三鋝，用致茲人。曾遘誨于氐[mask]：“[mask] [mask]舍酚矢五秉。”曰：“必當俾處厥邑，田[mask]田。”氐則俾復命曰：“諾。”昔饉歲，匡眾厥臣廿夫，寇曾禾十秭，以匡季告東宮，東宮迺曰：求乃人，乃弗得，汝匡罰大，匡遘稽首，于曾用五田，用眾一夫，曰噬，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，糧余。”曾又以匡季告東宮，曾曰：弋唯朕[mask]僦，東宮迺曰：“僦曾禾十秭，遣十秭，為廿秭，[mask]來歲弗僦，則付卅秭。”遘又即曾：用田二，又臣[mask]，凡用即曾田七田，人五夫，曾覓匡卅秭。

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

We first removed *Hu Ding* from the BIRD corpus to prevent data leakage. Then, we used a state-of-the-art SikuRoBERTa-based restoration model with greedy decoding (Lazar et al., 2021) to predict the missing parts of the inscription. The predictions are compared to gold-standard characters in Table 5.

Compared to the 22 characters restored by paleographic experts (Huang, 2022), our model achieved an Exact@1 of 40.91% (9/22), Exact@5 of 63.64% (14/22), and Exact@10 of 77.27% (17/22). In addition, it generated plausible completions for seven characters that remain undeciphered by experts (Table 6).

Mask Position	Pred@1	Top5
4	鑾	鑾旂馬舄車
8	伯	伯室守圃大
15	内	内外之邑大
16	于	于杜喬柞訊
17	則	則毋弗不勿
18	不	不帛毋勿弗
28	若	若其弋汝余

Table 6: Undeciphered characters and their predicted completions by our model.

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 categorized subset. These texts provide broad coverage of syntactic and lexical patterns closely aligned with inscriptional Chinese.

C Model Architecture

Our models consist of a transformer encoder domain-adaptive pre-trained on a large corpus of Pre-Qin Chinese texts, and task-adaptive pretrained on BIRD, extended with special tokens for unseen glyphs.

The training objective is a masked language modeling (MLM) loss:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i=1}^{N} \log P(y_i \mid X_i; \theta),$$

where y_i represents the target glyph, X_i the input context, and θ the model parameters.

To encourage consistent representations across allographs, we further integrate a glyph-net (GN) regularization term. For each variant group G, let $\mathcal{T}(G)$ denote the token set belonging to the same graphemic cluster, and let $\hat{p}_i$ be the predicted distribution at a masked position i. The GN loss is:

$$\mathcal{L}_{\text{GN}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \frac{1}{|\mathcal{T}(G_i)|} \sum_{t \in \mathcal{T}(G_i)} \log \hat{p}_i(t),$$

where $\mathcal{M}$ is the set of masked positions and G_i the variant group of the gold glyph at position i.

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

Table 7: Subset of Pre-Qin texts included in the DAPT corpus.

D Training Setup

We consider four adaptation schedules: (i) no adaptation, (ii) domain-adaptive pretraining (DAPT), (iii) task-adaptive pretraining (TAPT), and (iv) a two-stage pipeline of DAPT followed by TAPT. To disentangle the effect of structural priors, each schedule is further combined with two orthogonal mechanisms: a glyph net alignment and a group-biased masking bias. Three baseline models are all augmented with UNK placeholders and extended to cover unseen glyph characters. During DAPT, we freeze the bottom six transformer layers and train for ten epochs on a large Pre-Qin corpus. TAPT resumes with all layers unfrozen and adapts to inscriptional data. The two stages are interpolated with a weighting parameter λ that balances the contribution of DAPT and TAPT losses.

E Hyper-parameters

We found the best hyperparameters for each model during the search via WandB (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.

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, glyph-aware supervision, and biasing. Restoration accuracy is summarized in Table 9, representation cohesion and separation are analyzed in Table 10, and dating performance is reported in Table 11.

Model	Scenario	E@1 ↑	E@5 ↑	E@10 ↑	G@1 ↑	G@5 ↑	G@10 ↑
SIKUROBERTA	Baseline	23.59	37.73	43.95	24.37	39.47	45.75
	DAPT_only	25.95	42.31	49.37	25.26	43.17	51.15
	TAPT_Bias	48.25	62.60	67.55	54.39	67.84	73.07
	TAPT_GN	**49.47**	**65.20**	**70.15**	54.32	68.05	**73.07**
	TAPT_GN_Bias	49.24	63.83	68.58	**55.40**	**68.82**	72.86
	TAPT_from_DAPT	48.50	63.59	68.47	53.52	68.12	72.89
	TAPT_only	48.83	63.85	68.41	53.87	68.12	72.33
MBERT	Baseline	11.22	22.38	28.20	9.30	20.51	26.69
	DAPT_only	14.78	28.26	35.30	13.87	27.84	35.50
	TAPT_Bias	42.73	57.16	62.22	46.35	61.71	66.53
	TAPT_GN	**43.55**	**58.57**	**63.71**	**46.93**	61.28	65.92
	TAPT_GN_Bias	42.42	58.30	63.52	46.62	**61.77**	**66.46**
	TAPT_from_DAPT	43.11	57.37	62.34	46.38	60.68	65.68
	TAPT_only	42.73	57.03	61.27	46.53	60.62	64.77
XLM-BASE	Baseline	12.24	19.48	23.43	11.16	18.72	22.77
	DAPT_only	16.07	27.89	33.68	15.05	26.99	33.18
	TAPT_Bias	43.16	57.24	62.20	**45.38**	59.78	64.35
	TAPT_GN	**43.51**	**58.35**	**62.94**	44.28	59.49	64.03
	TAPT_GN_Bias	42.91	58.25	62.63	**45.38**	**60.78**	**65.10**
	TAPT_from_DAPT	43.38	58.27	62.88	44.66	59.46	63.87
	TAPT_only	42.35	55.69	60.22	43.37	56.80	61.43
XLM-LARGE	Baseline	14.02	22.50	26.53	13.20	20.79	25.66
	DAPT_only	17.80	32.11	38.19	16.64	31.20	38.44
	TAPT_Bias	45.31	59.51	64.00	**47.88**	**61.53**	65.59
	TAPT_GN	**45.64**	**60.92**	**64.91**	47.16	61.17	65.36
	TAPT_GN_Bias	45.35	59.78	64.47	47.58	60.85	**65.65**
	TAPT_from_DAPT	45.60	60.01	64.78	47.13	60.40	64.94
	TAPT_only	43.49	57.69	62.08	44.21	58.35	62.24

Table 9: Restoration results under different adaptation schedules across four pretrained models. E@k denotes Exact@k and G@k denotes Group@k.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 10: Representation analysis of variant 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 variant structure. 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	TAPT_GN_Bias	0.852	0.767	0.849	0.539	0.678	0.635
	DAPT_only	0.833	0.728	0.836	**0.552**	**0.684**	0.630
	TAPT_GN	0.840	0.698	0.842	0.542	**0.684**	**0.638**
	TAPT_Bias	**0.864**	**0.778**	0.842	0.543	0.671	0.629
	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
MBERT	TAPT_GN_Bias	0.815	0.642	0.829	**0.547**	0.651	0.601
	TAPT_GN	**0.846**	0.745	0.822	0.534	0.618	0.575
	Baseline	0.809	0.672	0.822	0.515	0.638	0.583
	TAPT_Bias	**0.846**	0.748	0.822	0.531	0.638	0.586
	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**
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_GN	0.765	0.429	0.803	0.433	**0.632**	0.569
	TAPT_Bias	**0.790**	**0.503**	**0.809**	**0.513**	0.625	0.573
	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_Bias	0.784	0.481	**0.809**	0.502	0.605	0.556
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_GN	0.821	0.705	0.809	0.512	0.572	0.534
	TAPT_Bias	**0.840**	**0.746**	0.816	0.531	0.651	0.630
	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_Bias	**0.840**	0.701	0.836	0.526	0.612	0.561

Table 11: 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.