

Language Coverage Is Not Variety Coverage: Lexical Underrepresentation of Classical Chinese in Multilingual Tokenizers

Lifan Deng

Rixin College, Tsinghua University

Department of Computer Science and Technology, Tsinghua University

dlf24@mails.tsinghua.edu.cn

Abstract

Multilingual models are commonly described as supporting Chinese, but language-level coverage may conceal disparities between historical and modern varieties. We audit five public or widely deployed tokenizer families on Classical Chinese prose and Modern Chinese newspaper text. XLM-R, a canonical multilingual representation model, requires 1.070 tokens per Classical character versus 0.688 for Modern Chinese. More importantly, only 17.4% of annotated two-character Classical words receive a single XLM-R token, compared with 66.4% of Modern words. After controlling for word length and corpus frequency, the odds of atomic coverage are 0.158 of the Modern odds. Four general-purpose LLM tokenizers reproduce the same asymmetry. Exact sentence-length matching, frequency controls, Traditional–Simplified pairs, and leave-one-source-out tests rule out several simple artifacts. Multilingual tokenizers can therefore cover a language’s script while underrepresenting the lexical units of its historical varieties.

1 Introduction

Multilingual models are evaluated primarily across named languages. This convention treats a language as covered when its script can be encoded and its modern benchmark variety appears in pre-training or evaluation. Yet a shared script does not imply shared lexical representation. Historical varieties may remain technically encodable while receiving far fewer useful vocabulary units than the modern standard.

Classical and Modern Chinese provide a controlled test. They largely share a character inventory, especially after script conversion, but differ in lexical distributions and grammatical conventions. A tokenizer can therefore appear to support “Chinese” while allocating its cross-character tokens mainly to Modern Chinese. The resulting Classical

text is not necessarily segmented *incorrectly*: Classical Chinese contains many monosyllabic words, and word boundaries are linguistically debatable. The stronger and testable question is narrower:

Do attested multi-character words in Classical Chinese receive systematically lower single-token lexical coverage than comparable Modern Chinese words?

We answer this question with a two-layer intrinsic audit. A large segmentation-free layer measures encoding cost for 5,000 Classical-prose and 10,000 Modern-newspaper sentences. A smaller reference layer measures whether annotated multi-character words coincide with tokenizer spans. The primary model is XLM-R (Conneau et al., 2020), whose shared SentencePiece vocabulary was trained for multilingual representation learning. Four current general-purpose LLM tokenizers test whether the result extends beyond one architecture.

Across all five tokenizers, Classical prose has both higher encoding cost and substantially lower lexical atomicity. For XLM-R, two-character Word Atomicity Coverage (WAC) falls from 66.4% in Modern Chinese to 17.4% in Classical Chinese; the frequency- and length-adjusted odds ratio is 0.158. Our contribution is thus not another observation that “Classical Chinese is split more.” It is evidence that *language coverage is not variety coverage*: multilingual vocabularies systematically underprovide atomic representations for established lexical units of a historical variety.

2 Related Work

Subword vocabularies optimize frequency and compression rather than linguistic boundaries (Senrich et al., 2016; Kudo and Richardson, 2018). Cross-lingual audits show that vocabulary allocation changes token cost and monolingual performance across languages (Rust et al., 2021; Petrov et al., 2023), while recent work connects tokenization to dialectal representation bias (Kanjirangat

et al., 2025). XLM-R itself exposes a capacity trade-off when one vocabulary and model cover many languages (Conneau et al., 2020). We move the unit of analysis from named languages to historical varieties within one writing system.

Classical Chinese word segmentation remains resource-intensive and standard-dependent (Huang and Wu, 2018; Ke, 2024). This uncertainty motivates our metric choice. We do not score tokenizer output against a complete gold segmentation. Instead, we ask only whether explicitly annotated multi-character words are available as atomic vocabulary units. Poetry tokenization has also received dedicated corpus and modeling work (Liu et al., 2022; Pan et al., 2023); because rhythm and line length confound historical period, poetry appears only as an appendix stress test.

3 Data and Method

Corpora. The main comparison contains 5,000 punctuation-free natural sentences of Classical prose and 10,000 Modern Chinese sentences from nine edited newspapers, including *People’s Daily*. The newspaper collection covers reporting and broader articles on social life. Its edited, standardized usage offers a naturally occurring record of contemporary written Chinese rather than a narrowly constructed benchmark. We impose no sentence-length cap. Classical text is represented in aligned Traditional and Simplified forms; Simplified Classical prose is the primary condition, so script cannot by itself create the historical–modern contrast.

Layer A therefore contains 15,000 base sentences and 20,000 physical forms for the primary prose comparison. Layer B contains 500 Classical-prose and 500 Modern sentences exactly matched by character length. Classical prose follows a conservative word guideline: stable lexicalized expressions, proper names, and official titles may be merged, while ordinary productive combinations remain separate. Modern segmentation starts from a coarse Chinese segmenter and is checked for exact reconstruction. This is a checked single-annotator reference, not a definitive gold standard. The Simplified comparison contains 2,067 annotated multi-character word instances. We additionally evaluate 500 Classical poetry lines as a stress test only (Appendix A).

Tokenizers. We evaluate XLM-R-base (250,002 vocabulary entries), GPT-4o’s o200k_base,

Tokenizer	Raw TPC		C–M gap	
	Classical	Modern	Matched	Adjusted
XLM-R	1.070	0.688	0.275	0.170
GPT-4o	1.083	0.775	0.273	0.172
Qwen2.5	0.924	0.597	0.302	0.217
LLaMA3	1.104	0.821	0.243	0.152
DeepSeek	0.891	0.552	0.311	0.222

Table 1: Layer A on Simplified text. C–M is Classical minus Modern. Matched gaps use exact-length pairs; adjusted gaps include all controls. Every reported TPC interval excludes zero.

Qwen2.5, LLaMA3, and DeepSeek-V3. No special tokens or normalization are added. Offset mappings determine token spans; GPT byte offsets are projected to character offsets. Character, Chinese BERT, and Chinese RoBERTa baselines are diagnostic and reported only in Appendix C.

Metrics. TPC is the number of tokenizer tokens divided by the number of input characters. SUB-CHARFRAG is the proportion of characters covered by two or more tokens, separating missing cross-character units from byte/subcharacter fallback. For every reference word of length at least two, WAC equals one iff a single tokenizer span exactly matches the word. TOKENBURDEN counts tokenizer spans overlapping the word. WAC measures available lexical atomicity, *not* word-segmentation accuracy or downstream quality.

Controls. For Layer A, we first match Classical and Modern sentences exactly by character length without replacement (2,941 pairs). We then regress each metric on a Classical indicator, log sentence length, mean log character frequency, rare-character rate, mean log bigram frequency, and digit/Latin rates. Frequencies are computed from the evaluation corpus and are observable proxies, not claims about proprietary pretraining counts. Standard errors are clustered by source. We also omit each newspaper in turn. Layer B uses logistic regression for WAC and OLS for TOKENBURDEN, controlling for word length, log string frequency, and mean character frequency, with sentence-clustered errors. Finally, Traditional–Simplified pairs isolate orthographic effects.

4 Results

The historical variety costs more to encode. Table 1 shows a large raw TPC gap for every tokenizer. Exact length matching preserves gaps of

Tokenizer	Two-char. WAC		Adj. OR	Δ Burden
	Classical	Modern		
XML-R	0.174	0.664	0.158	+0.327
GPT-4o	0.112	0.517	0.197	+0.301
Qwen2.5	0.264	0.716	0.194	+0.422
LLaMA3	0.096	0.399	0.404	+0.258
DeepSeek	0.357	0.716	0.265	+0.398

Table 2: Layer B lexical coverage. Adj. OR is the adjusted odds ratio for all multi-character words; Δ TOKENBURDEN is the adjusted Classical increment. All intervals exclude the null.

0.243–0.311. Adjustment for surface-frequency and form features reduces but does not remove them (0.152–0.222). XML-R requires 1.070 tokens per Classical character, compared with .688 for Modern Chinese, and retains a 0.170 adjusted gap (95% CI [0.159, 0.181]). Its adjusted SUBCHARFRAG difference is near zero, however. XML-R’s disparity is therefore driven mainly by scarce cross-character units, not failure to encode individual characters.

The gap targets attested multi-character words. The aggregate cost has a direct lexical interpretation (Table 2). XML-R represents 66.4% of Modern two-character words with one token, but only 17.4% of Classical words. After word length and frequency controls, the odds of atomic coverage are 0.158 of the Modern odds (95% CI for the log coefficient [−2.135, −1.552]); each Classical word overlaps 0.327 additional tokens. The four LLM tokenizers reproduce the same pattern: adjusted odds ratios range from 0.194 to 0.404, and adjusted burden increments from 0.258 to 0.422. Longer units strengthen rather than reverse the gap (Appendix B).

Script and source do not explain the result. Simplifying Classical text raises WAC and lowers burden for some tokenizers, but Simplified Classical prose remains far below Simplified Modern text. Across 1,218 paired Classical word positions, the Simplified-minus-Traditional WAC effect ranges from 0.003 to 0.071 for the four LLM tokenizers and 0.004 for XML-R. In Layer A, omitting each modern newspaper in turn leaves every raw TPC gap positive and narrowly distributed. The effect is therefore not driven by one publisher or by Traditional-character rarity.

5 What the Gap Is (and Is Not)

Atomic coverage is not segmentation accuracy.

A tokenizer is not required to reproduce linguistic word boundaries, and a character baseline is appropriate for many Classical spans. Our claim applies only to the annotated multi-character subset. WAC asks whether a vocabulary makes a reference lexical unit directly available; it does not penalize legitimate single-character words, nor does it assume that every syntactic or rhythmic group should be merged. This distinction matters for Classical Chinese, where treating all near-character encoding as error would build the desired conclusion into the evaluation.

Boundary crossing is a misleading success criterion.

Modern text actually has more tokens that cross a reference boundary. For XML-R, mean crossing-token rates are .076 in Modern news and .042 in Simplified Classical prose; Qwen shows .131 versus .065. This occurs because Modern text has many more cross-character tokens and hence more opportunities to over-merge. Conversely, a character baseline never crosses a boundary but has zero WAC for every multi-character word. The salient Classical failure mode is therefore underprovision of lexical atoms rather than frequent destructive merging.

Shared outcome, different mechanisms. XML-R’s adjusted SUBCHARFRAG gap is statistically indistinguishable from zero even though its adjusted TPC and WAC gaps are large. It generally encodes individual Classical characters cleanly but supplies too few multi-character vocabulary entries. GPT-4o and LLaMA3 additionally fragment more Classical characters internally, whereas Qwen and DeepSeek resemble XML-R’s missing-merge profile. A single fertility score would conflate these mechanisms. The combined metrics show why a script can be covered while a historical lexicon remains underrepresented.

6 Discussion and Conclusion

The audit supports one conclusion: multilingual tokenizers may cover a language while underrepresenting its historical varieties. “Chinese support” guarantees character encodability, not a vocabulary that reflects Classical Chinese lexical structure. XML-R is especially informative because it is explicitly multilingual, publicly reproducible,

and highly effective on modern cross-lingual benchmarks; nevertheless, its Classical two-character lexical coverage is less than one third of its Modern coverage.

This is an intrinsic representational asymmetry, not proof that character-level encoding is universally harmful. Composition from characters may recover meaning, and fine-grained units may benefit philological tasks. Conversely, higher WAC can reduce sequence length and make a lexical unit directly addressable by one embedding. The practical implication is evaluative: multilingual model audits should stratify not only by language, but also by historical, regional, and genre varieties. Adaptation work should then test whether vocabulary augmentation or continued pretraining reduces the measured gap and improves downstream Classical Chinese tasks.

Limitations

The primary comparison still confounds period with historical prose versus edited modern newspaper style; an aligned modern non-news corpus would improve causal attribution. Evaluation-corpus frequency is only a proxy for unavailable tokenizer training counts. Layer B is a checked single-annotator reference, and independent blind annotation is needed before calling it gold. The audit measures vocabulary availability, not contextual representations or downstream accuracy. Results are tokenizer-version-specific, and the five systems do not exhaust multilingual model designs.

Ethical Considerations

The study analyzes historical literary text and previously collected newspaper text. It neither profiles individuals nor infers sensitive attributes. Because redistribution rights for the sentence-level newspaper corpus are unclear, neither the evaluation text nor the testing code is planned for public release on GitHub. The paper instead reports tokenizer identifiers, preprocessing decisions, aggregate statistics, confidence intervals, and annotation rules. This limits exact third-party replication, which we state explicitly rather than distributing text without permission. Token efficiency should not be interpreted as the worth or linguistic quality of a variety.

References

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.

Shilei Huang and Jiangqin Wu. 2018. [A pragmatic approach for classical Chinese word segmentation](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. [Tokenization and representation biases in multilingual models on dialectal NLP tasks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yonghong Ke. 2024. [Construction of ancient Chinese word segmentation and part-of-speech corpus](#). In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics*, pages 819–829.

Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Yuchen Liu et al. 2022. [Introducing a large corpus of tokenized classical Chinese poems of tang and song dynasties](#). In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*.

Changzai Pan, Feiyue Li, and Ke Deng. 2023. [TopWORDS-Poetry: Simultaneous text segmentation and word discovery for classical Chinese poetry via bayesian inference](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language model tokenizers introduce unfairness between languages](#). *arXiv preprint arXiv:2305.15425*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. [How good is your tokenizer? on the monolingual performance of multilingual language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3118–3135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

A Poetry Stress Test

Poetry is excluded from the main historical–modern comparison because rhythm, fixed line length, and genre density confound period. It is retained only to test whether a constrained Classical genre intensifies the descriptive pattern.

Tokenizer	Two-char. WAC	Mean TOKENBURDEN
XLM-R	0.073	2.273
GPT-4o	0.034	2.304
Qwen2.5	0.165	1.860
LLaMA3	0.012	2.399
DeepSeek	0.242	1.825

Table 3: Simplified Classical poetry stress-test results.

B Longer Multi-character Words

In Classical prose, three-character WAC is zero for XLM-R, GPT-4o, Qwen2.5, and LLaMA3, and 0.034 for DeepSeek. Modern three-character words retain coverage of 0.092–0.362 across the five tokenizers. No Classical word of four or more characters is atomic under any tokenizer, while Modern coverage reaches 0.109. The adjusted regressions in Table 2 include word length, so different length distributions cannot by themselves generate the reported odds ratios.

C Diagnostic Baselines

A character baseline has $TPC = 1$, zero multi-character WAC, and no boundary-crossing tokens by construction. It therefore demonstrates why low boundary error is not evidence of good lexical coverage. Chinese BERT and Chinese RoBERTa are included as monolingual diagnostic baselines in the project-level aggregate analysis but not used to support the multilingual claim.

D Additional Robustness Details

Exact-length paired bootstrap intervals use 2,000 resamples. For XLM-R, the matched TPC gap is 0.275 with 95% CI [0.269, 0.282]. Leave-one-newspaper-out analysis never reverses the Classical–Modern difference for any tokenizer. Traditional–Simplified comparisons preserve sentence identity and annotated word positions. Removing boundary markers reconstructs every reference sentence exactly.

E Reference Segmentation Guideline

Reference boundaries use a half-width vertical bar. Single-character words remain separate by default. Stable lexicalized multi-character words, proper names, and official titles may be merged; ordinary productive syntactic combinations and rhythmic groupings are not automatically treated as words. Ambiguous analyses are retained separately. These choices deliberately make WAC a conservative coverage diagnostic rather than a claim that one segmentation theory is uniquely correct.