Was Poetry Graded Validly?: Text Mining *Shipin*, a Sixth-Century Chinese Work of Literary Criticism

Wenyi Shang[1], and Emily Xueyue Liu[1]

[1] School of Information Science & Learning Technologies, University of Missouri, Columbia, USA

Abstract

This paper applies text mining to investigate *Shipin* 詩品 (Poetry Gradings), a sixth-century Chinese work of literary criticism. Using a BERT model fine-tuned with masked language modeling on a classical Chinese poetry corpus, we generated embeddings for *Shipin*'s evaluative remarks on each poet and their own poetry corpora, and explored the relationship between these embeddings and the grades assigned to each poet by *Shipin* using PCA and machine learning classification. We found that both remarks and poetry provide some justification for the assigned grades, with remarks showing a much closer alignment. A poet's dynastic period and poetic origin influenced the grades they received in nuanced ways, reflecting *Shipin*'s preference for poetic styles. The results indicate that *Shipin* maintained an implicit but consistent standard in grading.

Keywords: literary criticism, classical Chinese poetry, text mining, BERT embeddings, classification

1 Introduction

The *Shipin* 詩品 (Poetry Gradings) of Zhong Rong 鍾嶸 (469?–518 C.E.) is the first systematic work of literary criticism on poetry in Chinese history. It bequeathed a profound legacy to later literary criticism, serving as the progenitor of the "remarks on poetry" (*shihua* 詩話) genre [28, p. 614], which became a central mode of poetic commentary in early modern East Asia, not only in China but also in Korea and Japan. Furthermore, in early medieval China, when the work was created, "literature was a political activity" since "literary skill was a prime factor in gaining official and social advancement" [12, p. 76]. Thus, *Shipin*'s influence extends beyond aesthetic and cultural debates to shape social and political life.

Shipin is organized into three volumes. Each volume begins with a preface outlining its literary theory, followed by critical entries on a total of 122 poets plus the anonymous "ancient poems" (*gushi* 古詩). For each poet, *Shipin* assigns one of three grades—upper, middle, or lower—to the person, offers a concise evaluative remark, and, for some prominent poets, locates them within a poetic lineage by identifying their poetic origins. As Wixted summarizes, this approach "characteriz[es] a poet in language that could (and generally did) refer both to (a) the personality or character of the writer and (b) the writings of the author—as well as sometimes (c) the response, that is, the feeling or impression that the poetry engendered in readers" [24, p. 275].

This threefold approach not only reverberated through classical Chinese literary criticism, but also imbues *Shipin* with such interpretive richness that it invites a multitude of often conflicting critical readings. For example, as to the work's aims, Wilhelm argues that "Chung Hung's [Zhong Rong's] basic position on problems of poetics does not appear indebted to any spiritual or doctrinal

Wenyi Shang, and Emily Xueyue Liu. "Was Poetry Graded Validly?: Text Mining *Shipin*, a Sixth-Century Chinese Work of Literary Criticism." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1067–1079. https://doi.org/10.63744/uNUzr0wn2VsQ.

tradition ... His interest is in poetics which he does not expound in religious or philosophical terms" [23, p. 117]. By contrast, Brooks sees Zhong Rong as a conservative critic reacting against the avant-garde school and declares, "The SP [*Shipin*], then, is essentially a Confucian remonstrance with its age" [1, p. 149]. Since each position is grounded in substantial textual evidence, debates like this remain unresolved. Consequently, later research has shifted toward moving beyond *Shipin*'s text to consider the broader cultural context of its era [27], reconstructing Zhong Rong's educational background to clarify the work's intertextual ties to classical works [5], or investigating micro-level questions, such as the possible interpolation of a sentence into the text [14].

Nonetheless, computational methods, bolstered by recent advancements in natural language processing, have the potential to reshape this landscape. In *Shipin*, the absence of "explicit criteria of judgment" and the pervasive "vague concrete approximations, poetically expressed, of traits perceived in a writer's work" [25, pp. 242–243] has driven scholars to divergent conclusions through close readings of individual cases. But what if we set aside close reading of individual passages, and apply text mining to examine *Shipin*'s criticism on the roughly 100 poets altogether?

Recent digital humanities research has applied text mining to premodern Chinese poetry, focusing primarily on poems from the Tang dynasty (618–907 C.E.) and later periods [21; 31]. For the early medieval era, scholars have used network analysis to examine the fifth-century text *Shishuo xinyu* 世說新語 (A New Account of the Tales of the World) [2; 15]. However, no early medieval poetry corpus has yet undergone text mining, and, more importantly, to date computational studies have concentrated exclusively on the literary works themselves, leaving the genre of literary criticism and its nuanced dialogue with those texts unexplored.

This paper therefore seeks to fill this gap in computational approaches to premodern Chinese literary criticism, and to engage directly with the scholarly debates surrounding *Shipin*. By integrating Transformer-based modeling with insights from traditional scholarship, it systematically analyzes every poet evaluated in *Shipin* to assess both the consistency of the work's grading standards that assign each poet to the upper, middle, or lower grades and the interplay between the evaluative remarks that each poet receives and their poetry corpus. Specifically, it asks:

(1) Do *Shipin*'s evaluative remarks justify the grades it assigns to each poet?

(2) Do the poets' own poetry corpora support their assigned grades?

(3) How do a poet's dynastic period and poetic origin influence the grade they receive?

2 Methods

The dataset for this paper comprises two primary sources. First, we draw on *Shipin jizhu* 詩品集註 (Collected Annotations on *Shipin*) [30], from which we manually transcribed every evaluative remark and recorded each evaluated poet's dynastic period and poetic origin. Second, we assembled the complete corpus of poems by those same poets from *Xianqin Han Wei Jin Nanbeichao shi* 先秦漢魏晉南北朝詩 (Anthology of Poetry from the Pre-Qin, Han, Wei, Jin, and Northern and Southern Dynasties) [10], which is the most comprehensive collection of extant Chinese poetry prior to the seventh century, downloading the full text from the "daizhige" GitHub repository (`https://github.com/garychowcmu/daizhigev20`) and then cleaning and preprocessing it.

Among the 122 poets evaluated in *Shipin*, 97 have poems included in *Xianqin Han Wei Jin Nanbeichao shi*, totaling 1,819 poems. To generate semantic embeddings of these poems and *Shipin*'s evaluative remarks on these poets that faithfully reflect the linguistic and stylistic patterns of pre-seventh-century Chinese texts, we fine-tuned a Transformer-based language model on a masked-language-modeling (MLM) task. To identify the optimal model for fine-tuning, we compared six representative Chinese BERT variants on our corpus, including two modern Chinese models [3; 6] and four classical Chinese models [4; 17; 20; 21]. Among these, `ethanyt/guwenbert-base` (hereinafter GuwenBERT) achieved the highest baseline MLM accuracy (28.21%; see Appendix

A for details on model names, corresponding pretraining domains, and achieved accuracies), so we selected it for our downstream experiments.

We then fine-tuned GuwenBERT using the AdamW optimizer [9] with a weight decay of 0.01, exploring a grid of learning rates, batch sizes, and masking probabilities. We used poems from *Xianqin Han Wei Jin Nanbeichao Shi* by contemporaries of the *Shipin*-evaluated poets who were not themselves evaluated in *Shipin* for fine-tuning, while poems by *Shipin*-evaluated poets formed an independent test. We excluded any poem exceeding BERT's 512-token limit, resulting in 4,458 of 4,496 poems (99%) in the fine-tuning dataset and 1,785 of 1,819 poems (98%) in the test set. The fine-tuning data were further split into training (80%) and validation (20%) subsets. To identify optimal hyperparameters, we performed a grid search over the following configurations: Learning Rates: {1e-5, 3e-5, 5e-5}; Batch Sizes: {16, 32}; Masked Probabilities: {0.10, 0.15}; Epochs: up to 100, employing early stopping (patience = 7, min_delta = 0.0001).

Ultimately, the optimal configuration (learning rate = 1e-5, batch size = 32, mask probability = 0.10, with the bottom six layers frozen) achieved a maximum validation MLM accuracy of 45.19% (see Appendix B for details on the procedure and logs of GuwenBERT hyperparameter tuning). On the independent test set, this model demonstrated robust generalization, achieving 35.22% accuracy in the MLM task, a 7.01% gain over the unfine-tuned baseline. The fine-tuned model was then used to generate embeddings for both the evaluative remarks in *Shipin* and the *Shipin*-evaluated poets' poetry corpora. For any poem exceeding BERT's 512-token limit, instead of simply excluding them, we split it into equal-length segments under this threshold and obtained a single poem embedding by mean-pooling its segment embeddings. Next, for each of the 97 poets evaluated in *Shipin* whose poems appear in *Xianqin Han Wei Jin Nanbeichao shi*, we computed: (1) An embedding of the evaluative remark assigned to that poet in *Shipin*; (2) An embedding of the poet's entire corpus by mean-pooling the embeddings of all their poems. These paired embeddings were then subjected to our subsequent analyses.

3 Findings and Discussions

Shipin's assignments of individual poets to the upper, middle, or lower grades have long been contested. Wang Shizhen 王士禛 (1634–1711 C.E.) derided several of these classifications as "positions are turned topsy-turvy, black and white are confused," and listed specific poets he believed Zhong Rong had misgraded [18, p. 4800]. In his preface, Zhong Rong himself conceded that "the allotment of a poet to one of my three grades should not be regarded as final" [29, p. 98]. Drawing on semantic embeddings of both evaluative remarks and poetry corpora, we undertook a comprehensive re-examination of these gradings.

We first analyzed the correlation between the similarity of evaluative remarks and that of poetry corpora for each pair of poets among the 97 poets in our dataset. To account for the inherent non-independence of these pairwise measures, we employed a Mantel permutation test [11], performing 1,000 random simultaneous permutations of the rows and columns of the evaluative-remarks distance matrix and recalculating Spearman's ρ and Pearson's r at each iteration to build null distributions. Although the observed Mantel correlation coefficients are positive (ρ = 0.105 for Spearman and r = 0.082 for Pearson), they do not reach statistical significance (p = 0.093 for Spearman and p = 0.145 for Pearson), indicating no statistical support for the hypothesis that poets evaluated similarly by *Shipin* composed correspondingly similar poetry.

These results seem to lend support to critiques of *Shipin*'s grading, suggesting that Zhong Rong did not employ a consistent standard whereby poets with similar poetry received similar remarks. To investigate this further, we performed principal component analysis (PCA) on the embeddings of evaluative remarks and poetry corpora using Scikit-Learn [13]. Each embedding was projected onto the first two principal components (PCs) and plotted in two-dimensional scatterplots colored by the poet's assigned grade. Clustering patterns emerge in both visualizations.

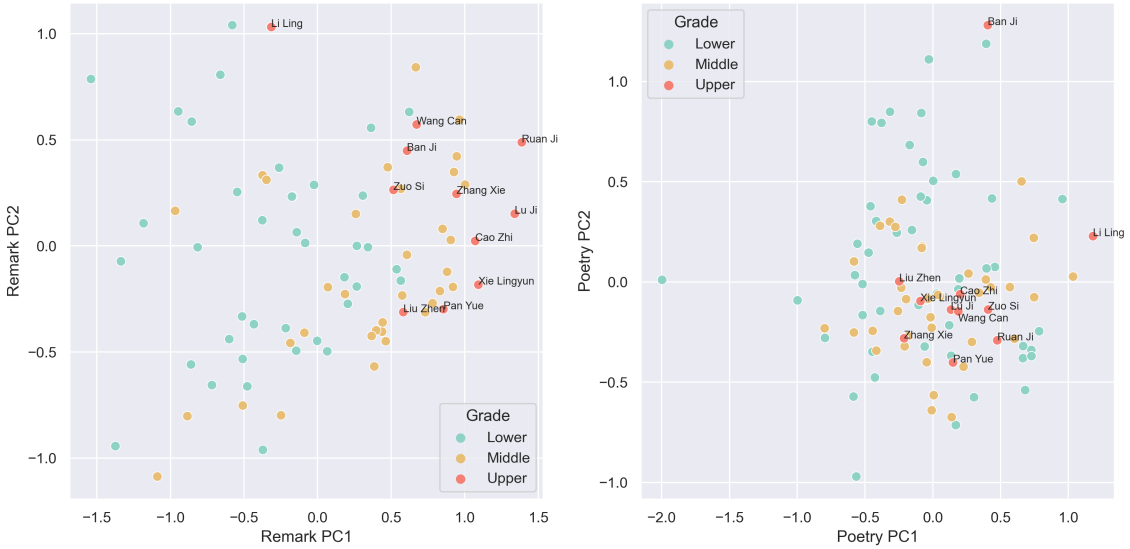

Figure 1: PCA of Remark and Poetry Embeddings, Colored by Grade.

In Figure 1, a clear gradient appears in the remark embeddings along the first PC: poets range from upper-grade on the right, through middle-grade in the center, to lower-grade on the left. Of the 11 labeled upper-grade poets, only Li Ling 李陵 falls below 0.5 on PC1, whereas only three of the 49 lower-grade poets exceed that threshold, indicating that *Shipin*'s gradings align closely with its evaluative discourse. The poetry embeddings reveal a subtler but still discernible pattern: upper-grade poets cluster tightly around the plot's center, while lower-grade poets are more widely dispersed toward the periphery. This central concentration suggests that the poets Zhong Rong grades highest tend to write in styles semantically similar to many others, implying they set the standards later poets emulate. This finding resonates with Wang Zhong's interpretation: drawing on Zhong Rong's evaluation of Cao Zhi 曹植, arguably the most highly praised poet in *Shipin*, he argues that the concept of "balance" (*zhong* 中) lies at the very heart of Zhong Rong's literary criticism [19]. Indeed, in our PCA of poetry embeddings, Cao Zhi (PC1 = 0.20, PC2 = -0.06) appears among the poets closest to the origin, embodying that balanced ideal.

Figure 2 shows the same PCA results colored by dynastic period rather than grade. In the remark-embedding space, no clear temporal ordering appears, confirming that *Shipin*'s evaluative discourse aligns with grading rather than chronology. By contrast, the poetry embeddings exhibit a diagonal gradient from the lower-right to the upper-left, mirroring dynastic succession from the earliest period covered in *Shipin*, the Han dynasty (202 B.C.–9 C.E., 25–220 C.E.), to the latest, the Liang dynasty (502–557 C.E.). This raises the question: how did this temporal stylistic evolution influence *Shipin*'s gradings? A simple inspection reveals Zhong Rong's chronological preference: of the six dynasties covered, the eleven upper-grade poets hail almost exclusively from the three earliest; only one comes from the third-latest Liu Song dynasty (420–479 C.E.); and none originate from the two latest dynasties. To further test the effect of dynastic period on grading in a comprehensive, data-driven way, we turned to machine learning classification.

We used leave-one-out cross-validation (LOOCV) to train two separate machine learning classifiers, one on remark embeddings and one on poetry embeddings, to evaluate how well each feature set predicts the grades *Shipin* assigns. During each LOOCV fold, we held out a single poet, trained a logistic regression classifier with balanced class weights using the Scikit-Learn [13] on the remaining 96 poets' embeddings and true grades, and then predicted the held-out poet's grade. After completing all folds, we compared the overall accuracies of the two classifiers and analyzed

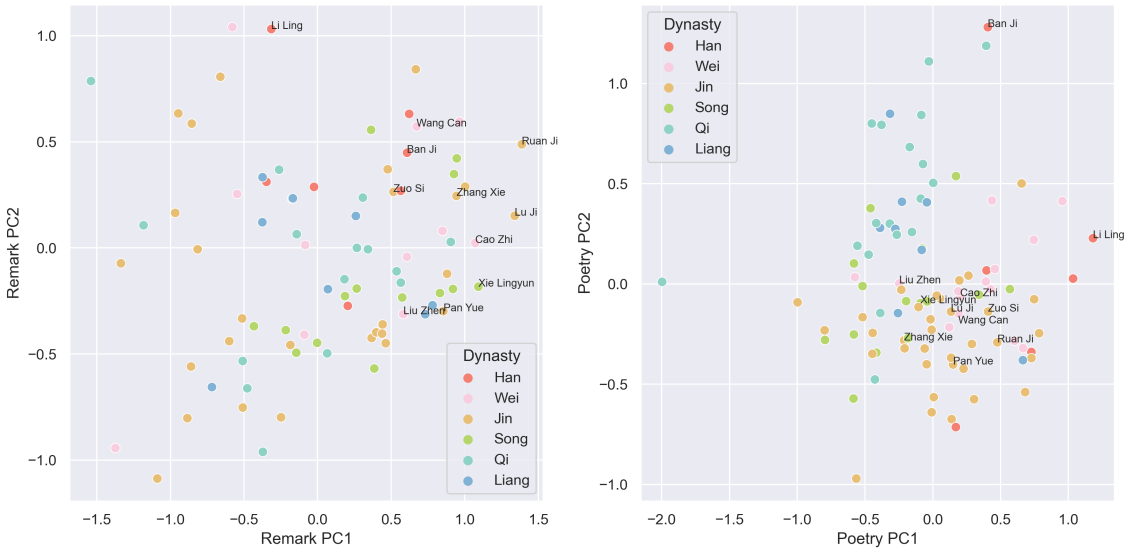

Figure 2: PCA of Remark and Poetry Embeddings, Colored by Dynastic Period.

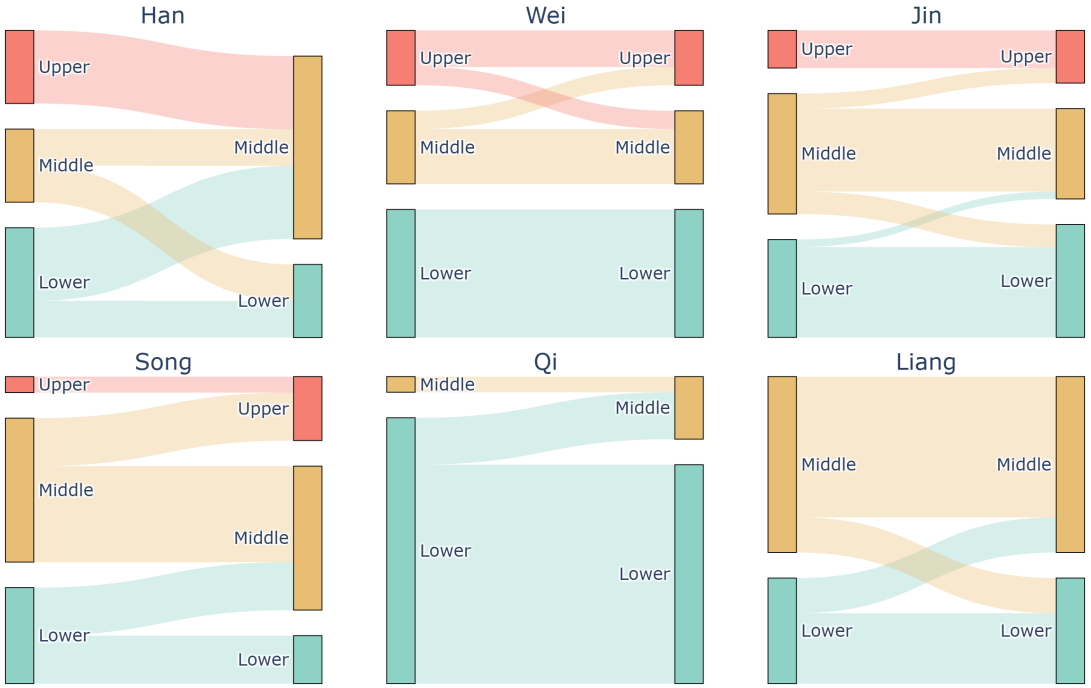

Figure 3: Remark-Based Classifier Predictions by Dynastic Period.

misclassifications in relation to each poet's dynastic period and poetic origin.

The logistic regression model trained on remark embeddings attains 75% accuracy, whereas the one trained on poetry embeddings reaches only 39%. Furthermore, unlike the poetry-based classifier, the remark-based classifier never misclassifies upper-grade poets as lower-grade, or vice versa. This disparity confirms our earlier finding that the grade *Shipin* assigned to each poet corresponds closely to its evaluative discourse, with only a subtler relation to the poet's poetic works.

In addition to overall accuracy, we next evaluated the performance of the two classifiers by dynastic period and poetic origin. The remark-based classifier exhibits an exceptionally low accu-

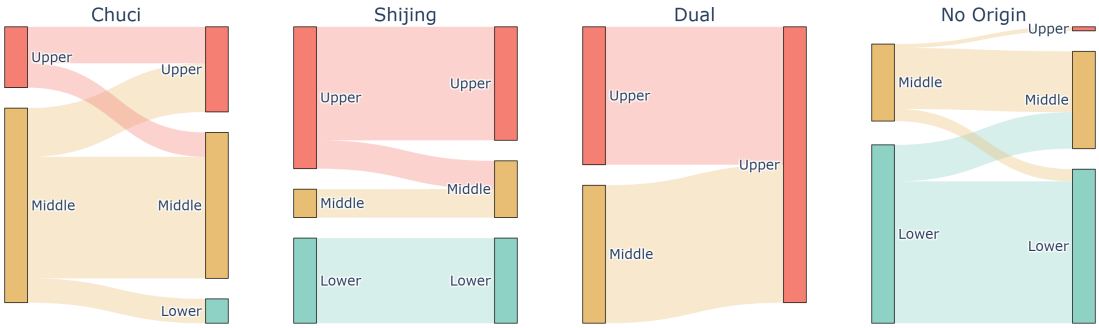

Figure 4: Remark-Based Classifier Predictions by Poetic Origin Group.

racy of 29% for poets from the earliest period covered in *Shipin*, the Han dynasty, whereas every other dynasty achieves at least 63%. As Figure 3 makes clear, Han poets are mostly misclassified downward: both upper-grade Han poets, Li Ling and Ban Ji 班姬, are labeled middle grade. Conversely, the classifier elevates six non-upper-grade poets to the upper grade, four of whom correspond exactly to the first four poets Wang Shizhen singled out as misgraded by Zhong Rong: Liu Kun 劉琨, Guo Pu 郭璞, Tao Qian 陶潛, and Bao Zhao 鮑照 [18, p. 4800]. This pattern echoes Wixted's argument that "the assignment of past poets to grosser-level gradings can additionally serve not only as a commendation or criticism of the writers themselves, but also as a commendation or criticism of those among one's contemporaries who emulate or advocate the emulation of their writings" [25, p. 245]. For Han poets, given the considerable temporal distance between their era and Zhong Rong's own, Zhong Rong likely relied on their eminent reputations and was unable to provide finely honed commendations, making their upper-grade status hard for the remark-based classifier to predict. By contrast, although Zhong Rong praised Liu Kun, Guo Pu, Tao Qian, and Bao Zhao in his evaluative remarks, his grade assignments may have been tempered by his views on contemporaries emulating their work, leading the classifier to misclassify them.

Particularly noteworthy among these four misclassified poets is Tao Qian, who was regarded as profoundly important in later literary history but was placed only in the middle grade by Zhong Rong. This placement has provoked such extensive criticism that "no part" of the work "has attracted more controversy than the entry on Tao Qian" [14, p. 553]. Our empirical results strongly reinforce Wixted's observation that "one can agree with the characterization while disagreeing with the grade assignment" [25, p. 245]: although Zhong Rong did give Tao Qian laudatory remarks that, as indicated by our remark-based classifier, suggest a higher standing—consistent with later critics' arguments—he nevertheless assigned him a middle grade. Moreover, we find that Tao Qian is not an isolated case: several other poets similarly received precise, positive evaluative remarks yet were assigned lower grades than their evaluations would warrant.

To examine how a poet's poetic origin influences their classification results based on the evaluative remarks they received, we traced each poet's origin to its root based on *Shipin*'s assignments. Among the 97 poets in our dataset, 21 have origins in *Chuci* 楚辭 (Verses of Chu), nine in *Shijing* 詩經 (Book of Songs), and two possess dual origins in both works. Classification accuracy is 89% for *Shijing*-origin poets, but only 62% for *Chuci*-origin poets. Poets belonging to neither group have a classification accuracy of 79%, falling between the two. Figure 4 reveals that *Chuci*-origin poets are misclassified in both directions: four upward and four downward. Thus, while Brooks claimed Zhong Rong preferred *Shijing*-origin poetry to *Chuci*-origin poetry [1, pp. 141–142], our empirical evidence suggests instead that Zhong Rong applied a more consistent evaluative standard to *Shijing*-origin works, as his evaluative remarks align more closely with the assigned grades for *Shijing*-origin poets than for *Chuci*-origin poets, making it easier for the classifier to produce accurate predictions.

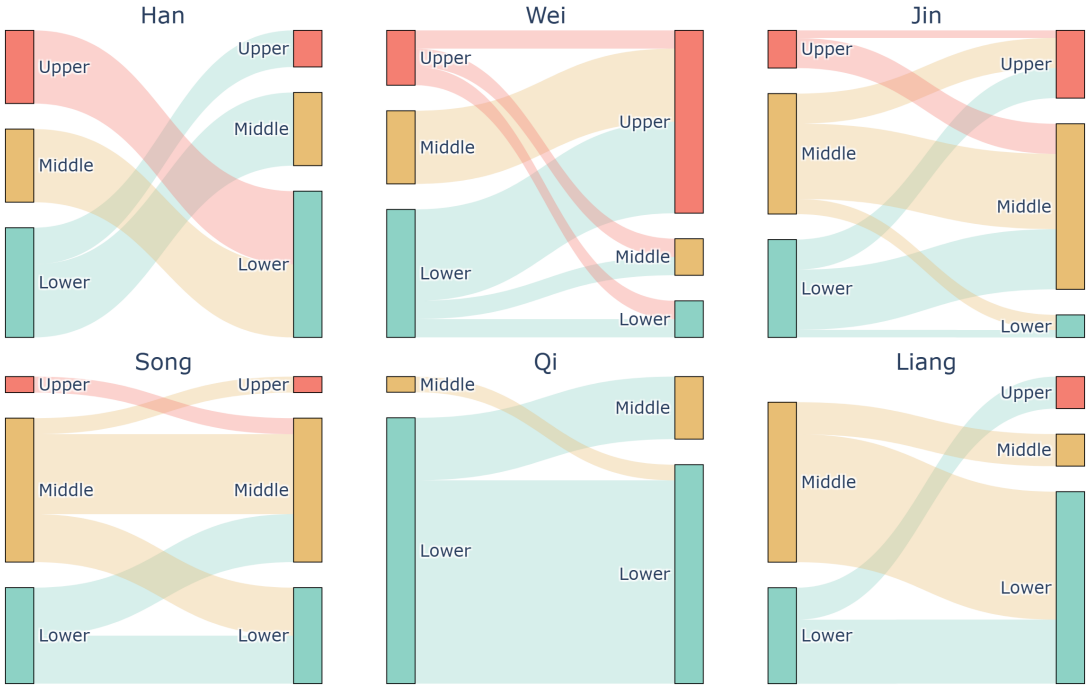

Figure 5: Poetry-Based Classifier Predictions by Dynastic Period.

For the poetry-based classifier, performance by dynastic period follows a pattern similar to that of the remark-based classifier. Poets from the two earliest dynasties, the Han dynasty and the Wei dynasty (220–266 C.E.), exhibit low accuracies of 0% and 14%, while every other dynasty achieves at least 35%. Specifically, as shown in Figure 5, among the 14 Wei-dynasty poets, ten are classified as upper-grade based on their poetry corpora, yet only three were actually assigned an upper grade. This upward misclassification supports Brooks's claim regarding Zhong Rong's preference for earlier poetic styles, who asserts that "[h]is advice to the poets of his generation was, in effect, to return to the early third century" [1, p. 145], referring precisely to the Wei dynasty.

In contrast, five of the six middle-grade poets from the two latest dynasties, the Southern Qi dynasty (479–502 C.E.) and the Liang dynasty (as noted earlier, no upper-grade poets hail from these dynasties)—namely Xie Tiao 謝朓, Jiang Yan 江淹, Fan Yun 范雲, Qiu Chi 丘遲, and Shen Yue 沈約—are downwardly misclassified as lower-grade based on their poetry corpora. However, four of them (all except Jiang Yan) are correctly classified as middle-grade based on the remarks they received. A similar pattern is observed for Xie Lingyun 謝靈運, an upper-grade poet from the Liu Song dynasty, who was downwardly misclassified as middle-grade based on his poetry but correctly classified as upper-grade based on his remarks. These results challenge Yeh and Walls's claim that Xie Lingyun's extensive use of allusions, which Zhong Rong principally opposed, was "admissible" to him as it was "complemented with a display of high literary genius" [26, p. 60]. Instead, they support Wixted's suspicion that "his placement of Hsieh Ling-yün [Xie Lingyun] in the upper grade of poets was in large measure owing to the latter's reputation in his age" [25, p. 245]. The contrast between Xie Lingyun's upper-grade-like remarks and middle-grade-like poetry —and similarly, the contrast between Xie Tiao, Fan Yun, Qiu Chi, and Shen Yue's middle-grade-like remarks and lower-grade-like poetry (these poets represent the avant-garde "Yongming poetic style" (*Yongming ti* 永明體), which Zhong Rong opposed)—indicates that Zhong Rong maintained an implicit but fairly consistent evaluative standard for poetry, while occasionally conceding to a poet's contemporary reputation in assigning grades and providing remarks.

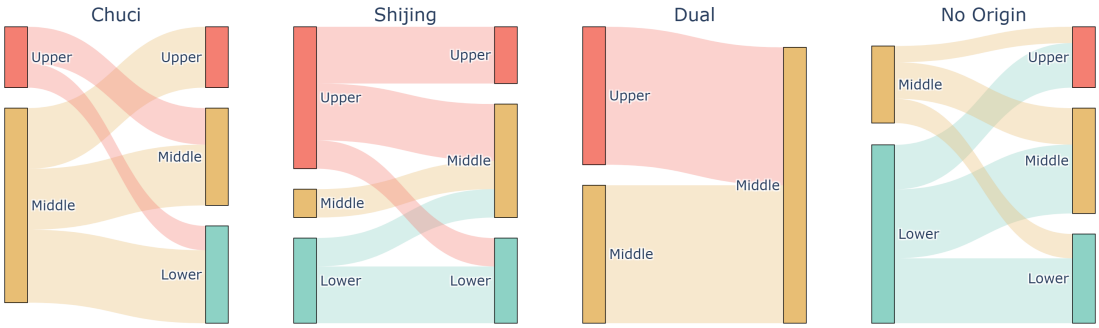

Figure 6: Poetry-Based Classifier Predictions by Poetic Origin Group.

The poetry-based classifier achieves an accuracy of 56% for *Shijing*-origin poets, 24% for *Chuci*-origin poets, and 40% for poets belonging to neither group. This pattern closely resembles that of the remark-based classifier as shown in Figure 4, which, as discussed earlier, indicates that *Shipin* applies a more consistent evaluative remark language when assigning grades to *Shijing*-origin poets. The similar pattern observed in the poetry-based classification in Figure 6 suggests a further insight: *Shipin* also maintains a more consistent standard regarding the poetic work itself for different grade levels among *Shijing*-origin poets. In other words, not only does the evaluative language align more consistently with grades for *Shijing* poets, but their poetry also more clearly supports these grading distinctions.

A final observation regarding the classification task is that only four poets are classified as upper-grade by both the remark-based and the poetry-based classifiers: Cao Zhi, Ruan Ji 阮籍, Liu Kun, and Ying Qu 應璩. Cao Zhi and Ruan Ji are highly canonical poets who were indeed assigned upper-grade status in *Shipin*. Liu Kun, as discussed earlier, is also canonical and was argued by Wang Shizhen to merit upper-grade classification. Ying Qu, however, is not widely regarded as significant within traditional literary history. Shen Deqian 沈德潛 (1673–1769 C.E.) strongly contested Zhong Rong's tracing of Tao Qian's poetic origin to Ying Qu [16, p. 182], and a later scholar has labeled such misassigned origins as Zhong Rong's "fundamental mistake" [8, p. 1659]. Nevertheless, Ying Qu's classification as upper-grade by both classifiers suggests that Zhong Rong highly valued him in his evaluative remarks, and that Ying Qu's poetic style aligns with Zhong Rong's standard for upper-grade poetry. Although Zhong Rong "fail[ed] to offer an objective theoretical explanation for his practice of source-tracing" [26, p. 48], he appears to have maintained an implicit standard in his grading, origin tracing, and evaluative remarks, according to which Ying Qu justifiably served as the poetic origin for a prominent figure like Tao Qian.

4 Conclusion

In summary, our findings reveal that both *Shipin*'s evaluative remarks and poets' own poetry corpora provide some justification for the grades assigned to each poet, with the evaluative remarks showing a much closer alignment. A poet's dynastic period strongly influences the assigned grade: Zhong Rong favored earlier poetic styles but was not always able to offer finely honed commendations for poets who lived long before him, and sometimes had to concede to a poet's contemporary reputation. As a result, earlier poets were often assigned grades higher than their evaluative remarks warrant but lower than their poetry suggests, whereas later poets frequently received grades lower than their remarks warrant but higher than their poetry suggests. Regarding poetic origins, both the remarks and poetry of *Shijing*-origin poets align more closely with their assigned grades than those of *Chuci*-origin poets, indicating that Zhong Rong maintained a more consistent evaluative standard for poets whose origin he traced back to *Shijing*.

We plan to pursue three directions in our next stage of research. First, we will expand this paper's focus by adding another dimension, the reception history of the poets evaluated in *Shipin*. By examining the relationship between our classification results and features such as the number of poems by each poet selected in anthologies across time, we aim to gain further insight into the place of *Shipin*'s evaluations within the broader scope of literary history. Second, we will refine our methods in several ways: by analyzing embeddings of individual poems in addition to the current mean-pooled embeddings of all poems by each poet; by adding an ablation study that masks overt praise or blame language in remarks to assess its impact on the remark-based classifier's performance; and by exploring additional BERT variants, such as ModernBERT [22]. Third, we will examine more interpretable features, such as lexicons, to investigate how the presence of specific terms in both the remarks and the poets' own poetry corpora predict upward or downward grades. We will also incorporate close reading to uncover additional insights into the stylistic nuances underlying our findings.

To conclude, while *Shipin* lacks "objectively defined standards" and embodies the "vagueness" considered "the greatest defect in the Chinese tradition of literary criticism" [26, p. 70], this paper showcases that, with the aid of computational methods, we can comprehensively examine the multifaceted interrelationships among assigned gradings, evaluative remarks, and poets' poetry corpora, revealing a broader picture that sheds new light on *Shipin*'s value in literary criticism. Despite occasional controversial grade assignments, Zhong Rong's evaluations exhibit consistent patterns far more intricate than a simple categorization of "which kinds of poets should be assigned to which grade," patterns that even he himself might not have been fully aware of. The evaluative judgments *Shipin* assigns to each poet are fundamentally relational, as "nobody has ever seen or imagined a value as an independent reality" [7, p. 97]. These underlying consistencies are most clearly unraveled within a relational space by simultaneously comparing the remarks and poetry of all *Shipin*-evaluated poets, an endeavor made possible only through computational methods.

Data and Code Availability Statement

All data and code necessary to reproduce the results presented in this paper are available at `https://github.com/wenyi-shang/Shipin`.

References

[1] Brooks, E. Bruce. "A Geometry of the *Shr Pǐn*". In: *Wen-lin: Studies in the Chinese Humanities*, ed. by Tse-tsung Chow. Madison: University of Wisconsin Press, 1968, pp. 121–150.

[2] Chen, Jack W., Borovsky, Zoe, Kawano, Yoh, and Chen, Ryan. "The *Shishuo xinyu* as Data Visualization". In: *Early Medieval China 20* (2014), pp. 23–59. DOI: 10.1179/1529910414Z.00000000013.

[3] Cui, Yiming, Che, Wanxiang, Liu, Ting, Qin, Bing, Wang, Shijin, and Hu, Guoping. "Revisiting Pre-Trained Models for Chinese Natural Language Processing". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, ed. by Trevor Cohn, Yulan He, and Yang Liu. Online, Nov. 2020, pp. 657–668. DOI: 10.18653/v1/2020.findings-emnlp.58.

[4] Ethan-yt. "guwenbert-base". Hugging Face Transformers. 2020. URL: https://huggingface.co/ethanyt/guwenbert-base.

[5] Führer, Bernhard. "Glimpses into Zhong Hong's Educational Background, with Remarks on Manifestations of the *Zhouyi* in His Writings". In: *Bulletin of the School of Oriental and African Studies 67*, no. 1 (2004), pp. 64–78.

[6] Google Research. “bert-base-chinese”. Hugging Face Transformers. 2018. URL: https://huggingface.co/bert-base-chinese.

[7] Hirsch, Jr., E. D. *The Aims of Interpretation*. Chicago and London: The University of Chicago Press, 1976.

[8] Hu, Yujin 胡玉縉. *Siku Quanshu Zongmu tiyao buzheng* 四庫全書總目提要補正 [Supplemental Corrections to the *Annotated Catalog of the Complete Library of the Four Treasuries*], ed. by Xinfu Wang 王欣夫. 1st. Shanghai: Zhonghua shuju Shanghai bianjisuo, 1964.

[9] Loshchilov, Ilya and Hutter, Frank. “Decoupled Weight Decay Regularization”. In: *7th International Conference on Learning Representations, ICLR 2019*. New Orleans, May 2019.

[10] Lu, Qinli 逯欽立. *Xianqin Han Wei Jin Nanbeichao Shi* 先秦漢魏晉南北朝詩 [Anthology of Poetry from the Pre-Qin, Han, Wei, Jin, and Northern and Southern Dynasties]. 1st. Beijing: Zhonghua shuju, 1983.

[11] Mantel, Nathan. “The Detection of Disease Clustering and a Generalized Regression Approach”. In: *Cancer Research 27*, no. 2_Part_1 (1967), pp. 209–220.

[12] Marney, John. *Liang Chien-Wen Ti*. Boston: Twayne Publishers, 1976.

[13] Pedregosa, Fabian et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research 12*, no. 85 (2011), pp. 2825–2830.

[14] Rusk, Bruce. “An Interpolation in Zhong Hong’s *Shipin*”. In: *Journal of the American Oriental Society 128*, no. 3 (2008), pp. 553–557.

[15] Shang, Wenyi and Sang, Zizhou. “Solidity in a Turbulent Flow: The Social Network of Aristocratic Families in the Eastern Jin Dynasty (317–420 C.E.)” In: *Journal of Historical Network Research 5*, no. 1 (2021), pp. 1–32. DOI: 10.25517/jhnr.v5i1.126.

[16] Shen, Deqian 沈德潛. *Gushi yuan* 古詩源 [The Origins of Ancient Poetry]. 1st. Beijing: Zhonghua shuju, 1963.

[17] Wang, Pengyu and Ren, Zhichen. “The Uncertainty-Based Retrieval Framework for Ancient Chinese CWS and POS”. In: *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, ed. by Rachele Sprugnoli and Marco Passarotti. Marseille, June 2022, pp. 164–168.

[18] Wang, Shizhen 王士禛. *Yuyang shihua* 漁洋詩話 [Yuyang’s Remarks on Poetry], ed. by Xiaowei Gong 宮曉衛. Vol. 6. Wang Shizhen quanji 王士禛全集 [The Complete Works of Wang Shizhen]. Jinan: Qi Lu shushe, 2007.

[19] Wang, Zhong 王忠. “Zhong Rong pin shi de biaozhun chidu 鍾嶸品詩的標準尺度 [Zhong Rong’s Standards and Criteria for Evaluating Poetry]”. In: *Guowen yuekan* 國文月刊 *66* (1948), pp. 25–28.

[20] Wang, Dongbo 王東波 and Liu, Chang 劉暢 and Zhu, Zihe 朱子赫 and Liu, Jiangfeng 劉江峰 and Hu, Haotian 胡昊天 and Shen, Si 沈思 and Li, Bin 李斌. “SikuBERT yu SikuRoBERTa: Mianxiang shuzi renwen de Siku Quanshu yuxunlian moxing goujian ji yingyong yanjiu SikuBERT与SikuRoBERTa：面向數字人文的《四庫全書》預訓練模型及應用研究 [Construction and Application of Pre-training Model of ‘Siku Quanshu’ Oriented to Digital Humanities]”. In: *Tushuguan luntan* 圖書館論壇 *42*, no. 6 (2022), pp. 31–43.

[21] Wang, Ziyao 王子堯 and Zhang, Jiandong 張建東. “Jiyu yuxunlian moxing de gudian shige fengge panding fangfa 基於預訓練模型的古典詩歌風格判定方法 [A Method to Judge the Style of Classical Poetry Based on Pre-trained Model]”. In: *Shuzi renwen* 數字人文 , no. 3 (2024), pp. 57–66.

[22] Warner, Benjamin et al. “Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference”. arXiv. 2024. URL: `https://arxiv.org/abs/2412.13663`.

[23] Wilhelm, Hellmut. “A Note on Chung Hung and His *Shih-p’in*”. In: *Wen-lin: Studies in the Chinese Humanities*, ed. by Tse-tsung Chow. Madison: University of Wisconsin Press, 1968, pp. 111–120.

[24] Wixted, John Timothy. “*Shi pin* 詩品 (Poetry Gradings)”. In: *Early Medieval Chinese Texts: A Bibliographical Guide*, ed. by Cynthia L. Chennault, Keith N. Knapp, Alan J. Berkowitz, and Albert E. Dien. Berkeley: Institute of East Asian Studies, University of California, Berkeley, 2015, pp. 275–288.

[25] Wixted, John Timothy. “The Nature of Evaluation in the *Shih-p’in* (Gradings of Poets) by Chung Hung (A.D. 469–518).” In: *Theories of the Arts in China*, ed. by Susan Bush and Christian Murck. Princeton: Princeton University Press, 1978, pp. 225–264.

[26] Yeh, Cha-Ying and Walls, Jan W. “Theory, Standards, and Practice of Criticizing Poetry in Chung Hung’s *Shih-p’in*”. In: *Studies in Chinese Poetry and Poetics*, ed. by Ronald C. Miao. Vol. 1. San Francisco: Chinese Materials Center, Inc., 1978, pp. 43–80.

[27] Zhang, Aidong. *Zhong Rong’s Shipin and the Aesthetic Awareness of the Six Dynasties*. PhD thesis. Toronto: University of Toronto, 1996.

[28] Zhang, Xuecheng 章學誠. *Wen shi tongyi jiaozhu* 文史通義校注 [Comprehensive Criticism of Literature and History, Collated and Annotated], ed. by Ying Ye 葉瑛. 1st. Beijing: Zhonghua shuju, 1985.

[29] Zhong, Rong. “Preface to *The Poets Systematically Graded*”. In: *Early Chinese Literary Criticism*, ed. by Siu-kit Wong. Trans. by Siu-kit Wong. Hong Kong: Joint Publishing Co., 1983, pp. 89–114.

[30] Zhong, Rong 鍾嶸. *Shipin jizhu* 詩品集註 [Collected Annotations on *Shipin*], ed. by Xu Cao 曹旭. 1st. Shanghai: Shanghai guji chubanshe, 1994.

[31] Zhu, Yuchen 諸雨辰 and Li, Shen 李紳. “Tang Song zhijian: Li Mengyang lüshi zhong de tong ti yi diao 唐宋之間：李夢陽律詩中的同題異調 [Between the Tang and the Song: The Same Title and Different Styles in Li Mengyang’s Metrical Poems]”. In: *Shuzi renwen* 數字人文 , no. 3 (2024), pp. 38–56.

A Baseline MLM Accuracy Across Chinese BERT Variants

Table 1 reports the MLM accuracy of six Chinese BERT variants evaluated on our poetry corpus.

B Hyperparameter Tuning of GuwenBERT

To fine-tune GuwenBERT effectively given the relatively small size of our poetry corpus, we explored several strategies to maximize generalization. First, while the conventional mask probability in BERT-style pre-training is 0.15, we hypothesized that a lower mask probability would allow the model to leverage more contextual information per input sequence, which is especially valuable for short texts like poems. We therefore compared performance under mask probabilities of 0.15 and 0.10, and found that 0.10 achieves consistently higher validation accuracy. Second, to

Model Name	Pretraining Domain	MLM Accuracy
google-bert/bert-base-chinese	Modern Chinese	12.87%
hfl/chinese-macbert-base	Modern Chinese (refined)	9.06%
ethanyt/guwenbert-base	Classical Chinese	**28.21%**
SIKU-BERT/sikubert	Classical Chinese	14.65%
Jihuai/bert-ancient-chinese	Classical Chinese	18.15%
qixun/bert-chinese-poem	Classical Chinese Poetry	16.39%

Table 1: Baseline MLM Accuracy (on Simplified and Normalized Corpus) across Six Chinese BERT Variants.

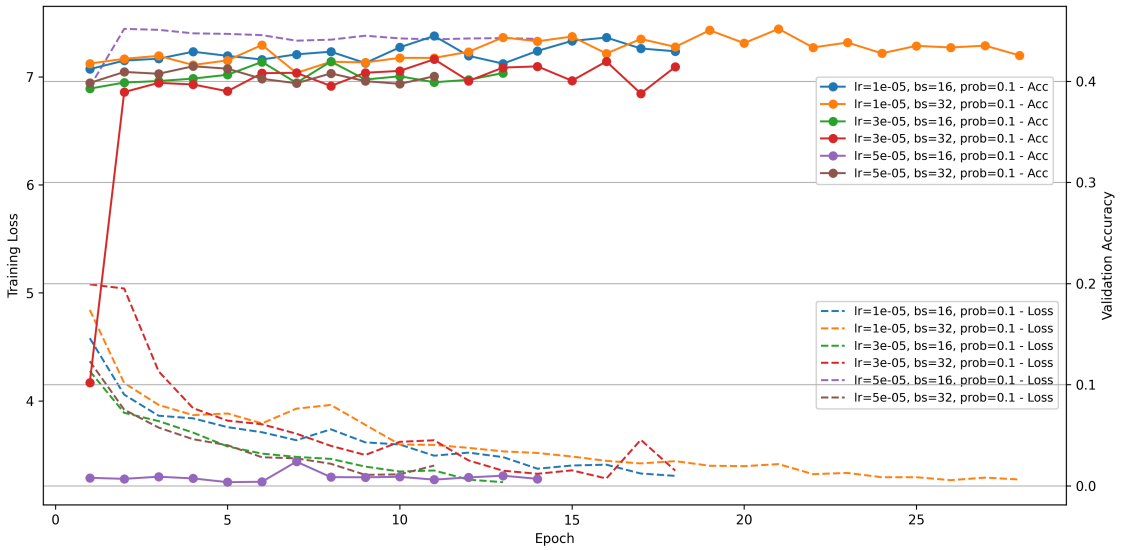

Figure 7: Training Loss and Validation Accuracy for Models with Mask Probability 0.15.

retain the general syntactic and semantic features captured during pre-training, we froze the bottom six layers of GuwenBERT and fine-tuned only the top layers. A summary of validation accuracy across all hyperparameter configurations is presented in Figures 7 and 8.

The best-performing configuration (learning rate = 1e-5, batch size = 32, mask probability = 0.10, with six frozen lower layers) achieved a maximum validation MLM accuracy of 45.19%. While the training loss continued to decrease throughout the optimization process, the validation accuracy plateaued after Epoch 21, indicating diminishing returns from further updates. We therefore selected the model checkpoint at Epoch 21 for later embedding generation and data analysis. The corresponding accuracy and loss curves are shown in Figure 9.

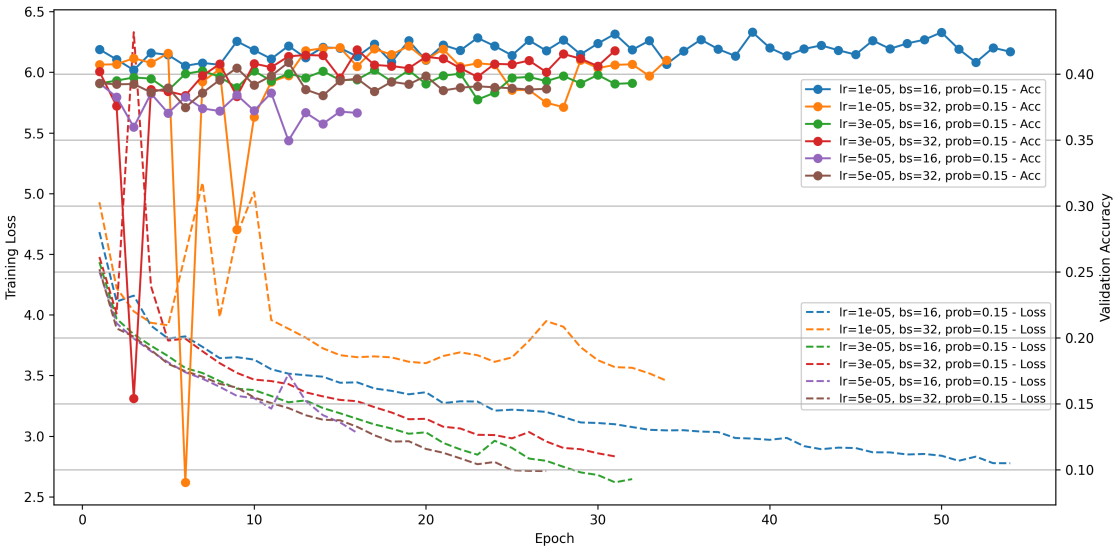

Figure 8: Training Loss and Validation Accuracy for Models with Mask Probability 0.10.

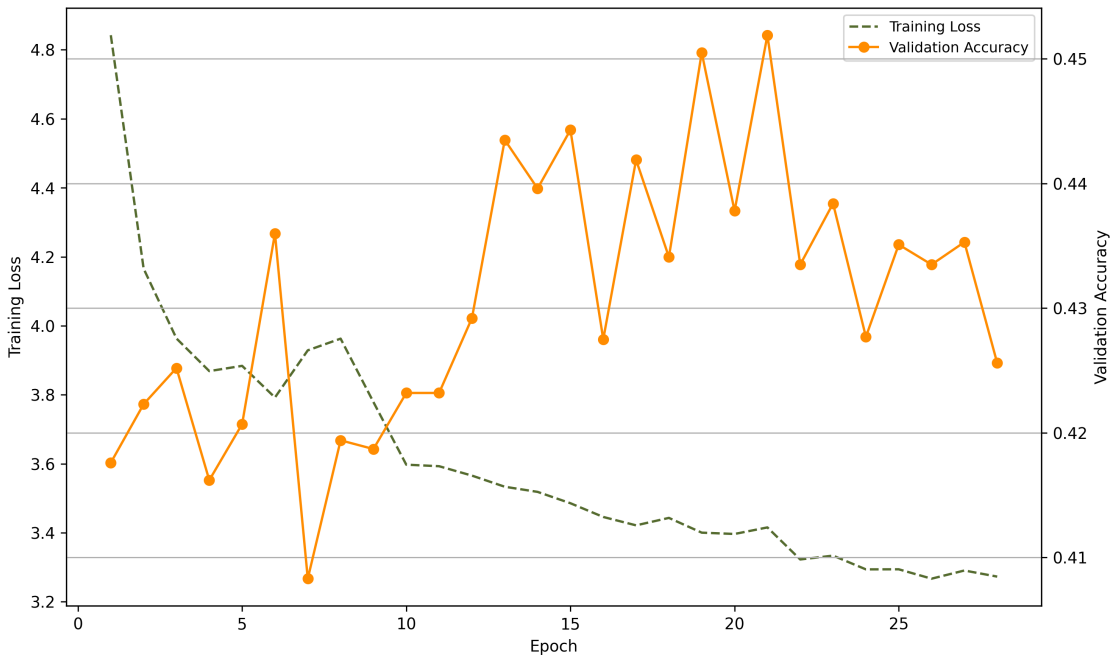

Figure 9: Training Loss and Validation Accuracy over 29 Epochs for the Model with Learning Rate 1e-5, Batch Size 32, and Mask Probability 0.10.