Innovation Humanities and Social Sciences Research, Volume 22, Issue 5, 2026. ISSN: 2949-1282
Proceedings of The 7th International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2026)

Research on Textual Prediction of Stele Lacunae Assisted by Top-5 Candidate Probabilities

Weixiao Hu[1], Hou Shu[2], Sihan Zhou[3], Baokun Yang[4], Shangge Li[5], Yujia Wang[6], Wei Song[7]

[1,2,3,4,5,6,7] *School of New Media, Beijing Institute of Graphic Communication, Beijing 102600, China*
[2]*Corresponding author. Email: shuhou@bigc. edu. cn*

ABSTRACT

The identification of missing characters in stele inscriptions is a critical factor for digital reconstruction of ancient texts. However, pure automated prediction has a limited accuracy for complicated contexts, and the single output of candidate 1 cannot provide satisfactory practical judgment. In order to address this problem, this paper constructs an evaluation corpus of inscription recovery based on 20 texts to compare the candidate generation ability between general Chinese models and ancient Chinese domain models using random masking. Experiment results indicate that GuwenBERT performs best among in epigraphic scenarios (Top1 accuracy: 50.95%, Top5 coverage: 71.73%). On this basis, this paper proposes a Top5 candidate probability assisted judgment scheme for missing characters. This method will display candidate characters, along with its probability of the bar chart superimposed on the original text. The results of a controlled user experiment prove that its effectiveness, i.e., when the information is probability, the accuracy of user judgment is increased from 41.50% to 57.50%, and the subjective confidence of the judgment score has increased from 3.04 to 3.60, That is, it can effectively help the identification of defective stele characters text through the display of candidate probabilities.

Keywords: *Stele inscriptions, Character prediction, Pre-trained language models, GuwenBERT.*

1. INTRODUCTION

The steles and rubbings have served as important mediums for traditional Chinese culture and they hold considerable historical significance and aesthetic value in the fields of calligraphy. Unfortunately, many rubbings have suffered from the loss of textual content due to natural erosion and damage from human sources. Thus, the important field of study in Digital Humanities today is based on the use of digitization as a successful method for predicting the missing characters.

In recent years, pre-trained language models (PLMs) have emerged as effective models for studying how a context can be modeled. Due to the masking language model feature, PLMs are able to predict the missing character accurately. There have been studies undertaken in order to compare the specialized ancient Chinese models with the normal Chinese models; however, results show very low Top-1 prediction accuracy for even the domain-adapted ones, thus limiting the acceptance of the single character.

Nonetheless, it must still be pointed out that even though Top-1 prediction has a low level of accuracy, a very good Top-5 predictions too can be achieved. Therefore, it is the challenge as to how to distribute the candidates to people in order for them to finalize their decision.

The question then becomes: is it acceptable to simply list the candidates? And do prediction probabilities have any relevant contribution to user decisions?

Basically, the target audience of this research includes all those who are interested in the stele culture, and have minimal form of knowledge of ancient Chinese.

The most significant contributions of this research are as follows:

- Creation of an evaluation corpus dedicated to the character prediction stage. Here, four PLMs were tested on the same evaluation subset for identifying finds the best model that should be used for experiments.
- Development of a system for assisting in judgments about the Top-5 candidates as well as their probability information by means of controlled validation.

2. RELATED RESEARCH

Previous studies regarding the restoration of absent characters in steles and rubbings can generally be and the two basic groups into which research can be categorized. The first group would be the image-based restoration techniques which make use the residual glyph data in order to fill in the character missing information. The second category would be the prediction based on context in which the missing data is drawn from a linguistic model developed from the previous theory. This study is purely about the second form of prediction and how to develop a linguistic model for the purposes of creating possible candidates based on the given unknown elements thereby providing assistance to the users in the event that the details within glyph are scant.

To the present day, researchers have utilized the use of Pre-trained Language Models (PLM) in order to gain further information through the linguistic models which learned through self-supervised learning via text linguistic model, including, the function of the Masked Language Model which focuses on the ability of the model to fill the missing words in the text, therefore it is very relevant when it comes to the prediction of missing characters. In the Chinese language sphere, the following models such as BERT-base-Chinese[1] and RoBERTA-wwm-ext[2] performed very well for the modern Chinese language but not that efficiently when it comes to the classical Chinese language as a result of large vocabulary, grammar, expression and lexical diversity that differentiates the two Chinese languages. Because of that issue, previous research studies have developed domain specific models for classical Chinese. GuwenBERT which was a language model trained on a large dene dataset of lexicon and even SikuBERT[3] that was trained on the Siku Quanshu data set. Both models have proven to perform quite effectively using this language model.

In the area of human-AI cooperative decision-making, conveying the uncertain predictions of AI models to users is a significant difficulty. Previous research has shown that effectively presenting a model's confidence information will help calibrate a user's trust in the model, thus leading to superior decision-making. For instance, in the area of natural language processing, presentation of Top-K candidates is an established way to convey uncertainty[4]. Input systems, machine translation post-editing systems, and optical character recognition error correction systems have all used the strategy of presenting candidates to users[5]. However, these systems present only the list of candidates, without any accompanying probabilities. So, users can only perceive indirectly the model's preferences for its choices based on the ordering of the candidates. In this paper, we will empirically explore the effectiveness of using probability display in the context of stele character prediction through a controlled experiment.

3. EVALUATION CORPUS AND CANDIDATE GENERATION

3.1 Construction of the Evaluation Corpus

This study evaluated the capability of language models to generate candidates to predict absent characters for stele restoration purposes. It built a special task-specific evaluation corpus for stele restoration, which was based on a raw corpus of 20 inscription texts comprising three main styles of writing: Kaishu, Lishu, and Wei-Bei. To ensure uniform length for the evaluation samples and to increase the sample size, the researchers used a sliding window technique, setting the length of the window to 80 characters with steps of 20 characters, thus producing 1,004 slices altogether. In conducting the missing character prediction experiments, exactly one character in an interior position was randomly chosen and masked out for each slice as a so-called context-based prediction. To address the issue of variability in character coverage for different tokenizers across models tested, only those slice samples were kept in which the masked character could be processed as one token. This left a final set of common evaluation samples venturing a total of 948 instances.

3.2 Candidate Generation Models and Evaluation Metrics

To compare the candidate generation capabilities of different language models in stele scenarios, four pre-trained models were selected, including two general Chinese models and two

ancient Chinese/ancient book models: BERT-base-Chinese, Chinese-RoBERTa-wwm-ext, SikuBERT, and GuwenBERT.

During the prediction process, given a text sequence $x = (x_1, x_2, \ldots, x_T)$ of length T, let the i-th position be the missing position. By replacing this position with a mask token, the context $x_{\setminus i} = (x_1, \ldots, x_{i-1}, [\text{MASK}], x_{i+1}, \ldots, x_T)$ is obtained.

The model outputs a logits vector z_i over the vocabulary for that position, and the prediction probability of character c is obtained via softmax:

$$P(c|x_{\setminus i}) = \text{softmax}(z_i)_c \quad (1)$$

The top k characters are selected in descending order of probability to form the candidate set $Top-k$:

$$C_k(x) = TopK(P(c|x_{\setminus i}), k) \quad (2)$$

Let N be the total number of samples in the test set, y_n be the ground truth for the n sample, and $C_k^{(n)}$ be the candidate set provided by the model. The $Top-k$ accuracy metric is defined as:

$$\text{Top-}k = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}(y_n \in C_k^{(n)}) \quad (3)$$

Some of the evaluation metrics are Top-1, Top-5 and Top-10. Top-1 accuracy reflects how well a model can give one unique output, while Top-5 and Top-10 indicate whether the right answer is included in the first 5 or 10 inputs. In this case, the first point selected Top-5 as the main indicator for visualization for two reasons: one is that Top-5 already provides a good number of candidates, fulfilling basic Assistant requirements for judgment; second is that Top-5 can be used instead of Top-10, which brings lower cognitive burden and the same complexity of interface, where Top-5 is better for real usage. However, Top-10 is still mentioned but will not be utilized in the next steps of design and usability.

3.3 Analysis of Model Results

The results of the four models on the common evaluation subset (N=948) are shown in “Table 1”.

Table 1. Candidate generation results of different models on the common evaluation subset

Model	Type	Top-1 (%)	Top-5 (%)	Top-10 (%)	N
BERT-base-Chinese	General Chinese	16. 77	29. 43	35. 34	948
Chinese-RoBERTa-wwm-ext	General Chinese	11. 29	20. 68	26. 16	948
SikuBERT	Ancient Books	23. 84	43. 35	48. 52	948
GuwenBERT	Ancient Chinese	50. 95	71. 73	78. 06	948

The findings show that general Chinese models have limitations in candidate generation for stele corpus, while models that have been trained on ancient books and ancient Chinese corpus perform much better. GuwenBERT is the only model that has outperformed the other models. A major outcome from this study is that GuwenBERT’s Top-5 coverage (71.73%) is much higher than Top-1 accuracy (50.95%). This shows in many cases where the model fails to rank the correct answer in the first place, it manages to include this answer among Top-5 candidates. Therefore, the focus turns from “whether the model can predict correctly” to the importance of techniques for helping users to make use of candidate information provided to them by the model.

4. CANDIDATE PROBABILITY PRESENTATION SCHEME AND USER EXPERIMENT

4.1 Design Goals and Scheme

In this research, the objectives for the format of its presentation have been condensed into three specific points. The first reason is to highlight the uncertainty of the model. The missings will use the Top-1 solution which gives the user the impression about the "confirmed answer" when in fact the real meaning of the solution is its differences in the probabilities of each candidates and how confident is the prediction model in outputting a certain answer. By doing this, people could use different

styles in making decisions, meaning that they would act differently while dealing with predictions with large-scale probabilities compared to those predictions with smaller ones based on the same model. The second reason is to simplify the decision process. The information about the nature of the stele writing makes this kind of reading more difficult than the reading of modern texts and therefore the template should exclude any kind of redundant information and make only the most necessary information for the decision-making process available. The third goal is to allow for the active decision-making process based on the context of the particular situation. Indeed, the key role of the information about the probabilities is to supplement but not to replace the user in the decision-making process. Thus, people should focus on the original situation in order to make a decision based on the context and not only use probabilities from the model.

Given the above-stated objectives, the authors use the following presentation template which includes using the title and subtitle, representing the core of the way how the user will see information in the form of "Punctuated Context and then Top-5 Candidates, Probability Bars, and Percentage based on these bars". The context section includes keeping the original punctuation and sentences in order to minimize the difficulty of reading. In the candidate section, the candidates will be sorted according to the model prediction level and will be represented in the form of the five possible options. The authors will use horizontal bars concerning the percentages in order for the users to perceive differences in the values of probabilities visually as well as to get supplementary information about the candidates more precisely. In this case, showing only the rank of the candidates will be less effective than showing the probabilities which would serve to illustrate different degrees of probabilities between them. In order to find out whether or not the use of the information about probabilities will be able to improve the decision process of users, two groups were developed:

- Control Group: The users will have only the Top-5 candidates and no information about the prediction model.
- Experimental Group: The users will be presented with the Top-5 Candidates as well as the respective probability information for each of them.

Both conditions have consistent content in terms of context material, number of candidates, order of candidates presented to users, and layout; the only possible difference between the two groups is whether or not the information about probabilities is present.

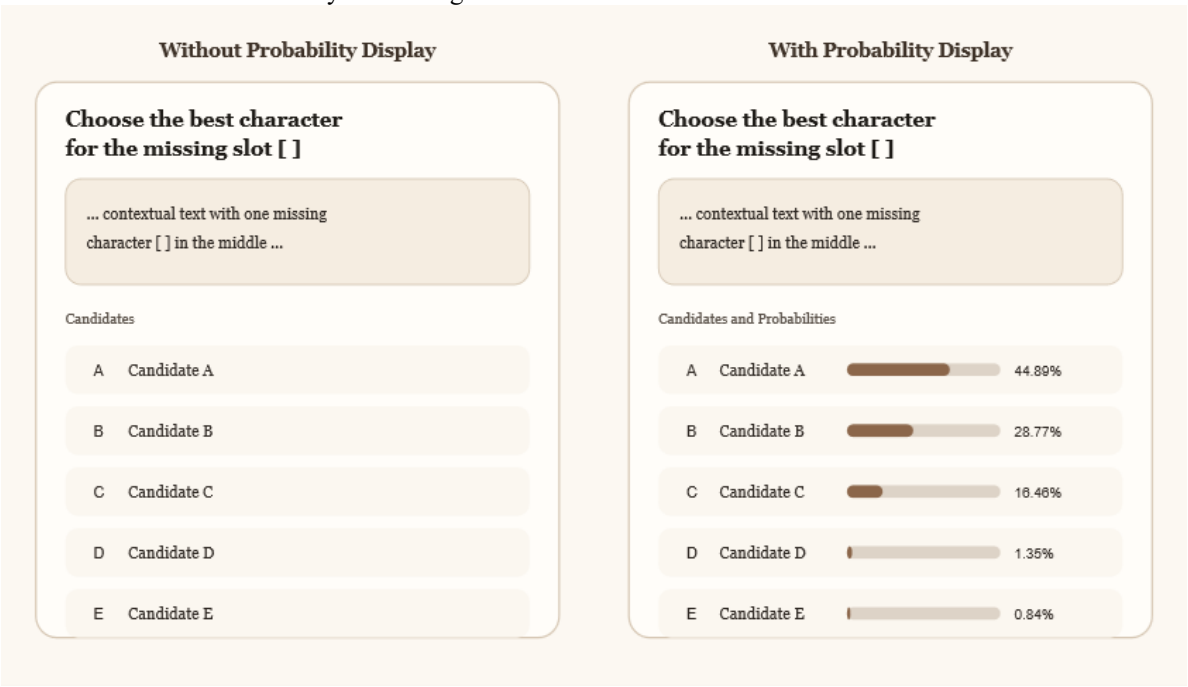

Figure 1 Top-5 candidate probability display prototype (Left: No probability information; Right: With probability information).

4.2 User Experimental Design

In this study, we conducted a within-subjects design where participants filled out either version A or B of a questionnaire composed of 20 questions. The participant's questionnaire contained 10 "control" and 10 "experimental" questions in random order. Specifically, the experimental questions were switched, so that if a question is a control question in version A, it is experimental in version B, and vice-versa. This design achieved the goal of having all questions be presented in both conditions to different people, which eliminated the potential for difficulty with the questions to interfere with our results.

In selecting the questions, we screened the evaluation corpus collected for use in this research project (Chapter 3) so that we retained only those items that had GuwenBERT top-five candidates that included the correct answer in that set of candidates. Therefore, we conducted stratified sampling to ensure a balanced distribution of question difficulty levels: eight high confidence items (top one probability over 70%), eight medium confidence items (top one probability between 30-70%), and four low confidence items (top one probability less than 30%), for a total of 20 queries to be created.

Regarding the participants, we selected a user group that was not limited to professional archaeologists or paleographers but rather included interested non-experts who have an understanding of ancient Chinese and interest in stele culture. A total of 50 subjects were recruited (25 in each version), whose educational backgrounds and experience reading ancient Chinese were obtained through the questionnaire.

All subjects read a chosen snippet of stele text punctuated in order to identify the most likely stele character from the choices A-E based upon the candidate information. Finally, all participants rated their confidence in their judgments on a scale from one to five.

4.3 Experimental Results

"Table 2" presents the judgment accuracy under different display conditions. In both Version A and Version B, the average accuracy under the probability-enabled condition was higher than that of the no-probability condition. After merging the 50 questionnaires, the overall average accuracy increased from 41. 50% to 57.50%, an improvement of 16.00 percentage points. These results indicate that, under the current experimental conditions, candidate probability presentation effectively enhances user judgment performance for missing stele characters.

Table 2. Judgment accuracy under different display conditions (%)

Questionnaire	N	No Probability Display	With Probability Display	Improvement
Version A	25	47. 00	56. 00	+9. 00
Version B	25	36. 00	59. 00	+23. 00
Total	50	41. 50	57. 50	+16. 00

"Table 3" shows the subjective confidence scores. The average confidence score rose from 3.04 to 3.60 under the probability-enabled condition, an increase of 0.56 points. Notably, the rise in confidence was accompanied by an actual improvement in accuracy, suggesting that participants indeed made more accurate judgments with the aid of probability information.

Table 3. Subjective confidence scores under different display conditions

Questionnaire	N	No Probability Display	With Probability Display	Improvement
Version A	25	3. 33	3. 54	+0. 21
Version B	25	2. 75	3. 67	+0. 92
Total	50	3. 04	3. 60	+0. 56

Based on participant responses after the study, the general consensus was that the probability information carried was beneficial for limiting choice options as well as for confirming prior thoughts and beliefs, and therefore, probability presentation played a favorable support role in most instances. Also, a handful of respondents admitted to having reservations when their own judgements did not concur with the model's high-probability recommendation. This is consistent with the outcomes of other research about "automation bias" [6] and indicates that future systems may benefit from designing a differentiated presentation of information in consideration of model confidence.

5. CONCLUSION

To address the challenge of producing outputs with low accuracy for finding missing characters in steles and rubbings via a fully automated approach, we propose a supplementary judgment technique founded on Top-5 candidate probabilities. Our study delineates the building of a specific evaluation corpus of 20 inscriptions and the comparative analysis of four pre-trained language models that led to the selection of GuwenBERT as the best candidate generation model utilizing Top-1 accuracy of 50.95% and Top-5 coverage of 71.73%. Based on these results, we established a presentation scheme utilizing Top-5 candidates with their predicted probabilities and validated it by user experiments. The findings indicate that the average judgment accuracy was enhanced from 41.50% to 57.50% (an increase of 16.00 percentage points), and the average subjective confidence rating elevated from 3.04 to 3.60 (an increase of 0.56 points) with the utilization of probabilities. In conclusion, the presentation of probabilities may benefit participants' use of model outputs to some extent, resulting in relatively superior outcomes concerning the identification of missing characters.

This study is limited by the small scale of the evaluation corpus and number of participants. Moreover, the investigation was solely limited to judgment tasks and did not encompass glyph fragments as well as images and results from OCR. Future work may increase the size of the evaluation corpus, include more diverse user groups, or utilize adaptive presentation methods customized according to model confidence. Further integration of probabilities concerning candidates and glyph generation modules may lead to the development of a wider and better stele restoration system.

ACKNOWLEDGMENTS

This study is supported by research programs of National-Level(and Beijing Municipal-level) College Student Innovation &Entrepreneurship Training Program of BIGC(Project Number: 202510015020,S202510015007X).

REFERENCES

[1] Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1, 4171–4186. https://doi. org/10. 18653/v1/N19-1423

[2] Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021:29, 3504–3514. https://doi. org/10. 1109/TASLP. 2021. 3124365

[3] Wang, D., Liu, C., Zhu, Z., Liu, J., Hu, H., Shen, S., & Li, B. Construction and application of pre-trained models of Siku Quanshu in orientation to digital humanities. Library Tribune, 2022:42(6), 31–43.

[4] Parasuraman, R., & Manzey, D. H. Complacency and bias in human use of automation: An attentional integration. Human Factors, 2010:52(3), 381–410. https://doi. org/10. 1177/0018720810376055

[5] He, S., Zhang, Z., & Chen, Y. To trust or not to trust: How different forms of confidence indicators affect trust and decision making in human-AI collaboration. International Journal of Human–Computer Interaction. 2023. https://doi. org/10. 1080/10447318. 2023. 2247514

[6] Buçinca, Z., Malte, B., & Gajos, K. Z.. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 2021:5(CSCW1), Article 188, 1–21. https://doi. org/10. 1145/3449287