SikuGPT: A Generative Pre-trained Model for Intelligent Information Processing of Ancient Texts from the Perspective of Digital Humanities

CHANG LIU, DONGBO WANG, ZHIXIAO ZHAO, DIE HU, MENGCHENG WU and HAI ZHANG
College of Information Management, Nanjing Agricultural University, Nanjing, China

LITAO LIN, JIANGFENG LIU
School of Information Management, Nanjing University, Nanjing, China

SI SHEN
Group of Science and Technology Full-text Knowledge Mining, School of Economics & Management, Nanjing University of Science and Technology, Nanjing, China

BIN LI
College of Liberal Art, Nanjing Normal University, Nanjing, China

LIANZHEN ZHAO
School of Foreign Languages, China Pharmaceutical University, Nanjing, China

CCS CONCEPTS • Human-centered computing → Human computer interaction (HCI) • Computing methodologies → Natural language generation • Applied computing → Arts and humanities

The rapid development of generative artificial intelligence has brought significant opportunities for the advancement of digital humanities research. Intelligent processing of ancient texts, as an essential part of digital humanities, is also undergoing a transformation in research methodologies in the wave of AIGC. The integration of generative pre-trained models with Chinese ancient texts, a vital carrier of Chinese culture, allows for deep mining of the content of these texts and provides services that make ancient texts more understandable and accessible to the general public. In this research, we propose a method that combines the most renowned Chinese anthology, the "Siku Quanshu," with generative pre-trained models. We developed the SikuGPT model, a generative model for ancient text processing tasks, based on GPT-type language models by continued pretraining. This model was tested on two typical tasks of ancient text processing: translation between classical and modern Chinese, and classification of ancient texts. The findings reveal that our model achieves advantages in understanding and generating scenarios of ancient texts. The capability of SikuGPT in processing traditional Chinese texts helps to promote the organization of ancient information and knowledge services, and advances the international dissemination of traditional Chinese culture.

Additional Keywords and Phrases: Generative Pre-trained Model, Siku Quanshu, Chinese ancient texts, Digital humanities research, Natural language processing

The authors acknowledge the National Social Science Foundation of China (Grant Numbers: (21&ZD331)) for financial support. We thank all the volunteers and all publications support and staff who wrote and provided helpful comments on previous versions of this document.
Authors' addresses: Chang Liu, Dongbo Wang, Zhixiao Zhao, Die Hu, Mengcheng Wu and Hai Zhang, College of Information Management, Nanjing Agricultural University, Nanjing, 210095, China; emails: 2023214005@stu.njau.edu.cn, db.wang@njau.edu.cn, zhaozhixiao@stu.njau.edu.cn, butterfly@stu.njau.edu.cn, wmc@stu.njau.edu.cn, 1033462760@qq.com; Litao Lin, Jiangfeng Liu, School of Information Management, Nanjing University, Nanjing, 210023, China; emails: litaolin@smail.nju.edu.cn, jfliu@smail.nju.edu.cn; Si Shen, Group of Science and Technology Full-text Knowledge Mining, School of Economics & Management, Nanjing University of Science and Technology, Nanjing 210094, China; emails: shensi@njust.edu.cn; Bin Li, College of Liberal Art, Nanjing Normal University, Nanjing 210097, China; emails: libin.njnu@gmail.com; Lianzhen Zhao, School of Foreign Languages, ChinaPharmaceutical University, Nanjing 211198, China; emails: lianzhen.zhao@cpu.edu.cn

ACM 1556-4711/2024/7-ART
https://doi.org/10.1145/3676969

1 INTRODUCTION

Nowadays generative AI has made remarkable achievements in various fields. The diffusion models that automatically generate images, cross-modal models that generate videos with one click, and the astounding ChatGPT all showcase the charm of AIGC (AI-Generated Content). The China Academy of Information and Communications Technology pointed out in its "White Paper on Artificial Intelligence-Generated Content (AIGC) " [White Paper on Artificial Intelligence-Generated Content (AIGC) (2022) -- China Information and Communication Academy, n.d.] that AIGC will become a unique information production method in the web3.0 era, following PGC (Professionally Generated Content) and UGC (User Generated Content). Large-scale pre-trained models have officially entered people's lives as productivity tools, bringing about tremendous changes to the industry.

Although there has been a considerable amount of research exploring AI technology for the intelligent processing of ancient Chinese texts [Lin & Wang. 2023], the tools employed in these studies are predominantly discriminative language models. Due to their inherent structure, discriminative models are primarily capable of supporting the annotation of grammatical and knowledge units within the texts, but they face challenges in achieving creative generation of content from the literature. The meeting of ancient texts and generative AI has injected new vitality into the intelligent processing of Chinese ancient texts. Using pre-trained models to generate texts can assist in the interlingual translation and intralingual translation of Chinese ancient texts and knowledge organization work. Currently, large language models represented by ChatGPT have been able to achieve barrier-free interaction with humans and can complete different natural language processing tasks according to human instructions. However, they may yield higher error rates when answering certain vertical field questions, because the training process of large language models focuses more on learning general knowledge than specialized knowledge. At the same time, using subject knowledge for fine-tuning or writing prompts for context learning is too costly, and domain-specific deployment in real scenarios remains very challenging. Previous research [Moradi et al. 2022] has indicated that in text processing in the field of biology, it is difficult for GPT3 to exceed the ability of small models in solving small sample tasks under the condition of the same corpus pre-training and fine-tuning. The processing of ancient Chinese texts is also a highly specialized task. On one hand, there have been significant changes in the meanings, morphology, and syntax of modern Chinese compared to ancient Chinese, processing and utilizing ancient texts requires researchers to possess a certain level of professional knowledge. On the other hand, compared to the vast and ever-increasing amount of modern Chinese language data, the amount of ancient text data is relatively scarce and no longer increasing. According to statistics from the "General Catalog of Chinese Ancient Books" platform [General Catalog of Chinese Ancient Books, n.d.] as of 2024, the total number of preservation records for Chinese ancient books collected by the project both domestically and overseas amounts to about 250,000 entries. Additionally, statistics from the Chinese Text Project platform [Sturgeon, n.d.]shows that the total number of characters in ancient resources that have been digitized and stored as text is only around 5 billion characters (including different versions of the same literature). The scarcity of resources and the difficulty of understanding ancient Chinese determine that high-quality ancient Chinese language corpora cannot be produced and obtained on a large scale and quickly. Therefore, the intelligent processing of ancient Chinese classics also falls under the category of NLP tasks in low-resource environments.

Using self-supervised learning methods to train language models with massive amounts of unlabeled data is an important means to alleviate the scarcity of labeled data. In this study, we develop a dedicated ancient texts processing model SikuGPT based on the best small pre-training model GPT2 for generative processing of ancient texts. We verified the performance of the generative model in two categories of tasks: text translation and text comprehension. In this article, we also released an open-source Ancient Chinese-Modern Chinese translation model fine-tuned with a bilingual parallel corpus, with the twofold aim of promoting research efficiency for scholars working in Chinese ancient texts and boosting the international dissemination of Chinese ancient culture.

2 RELATED RESEARCH

2.1 Generative Pre-trained Language Models

After going through statistical learning models and deep neural network models, NLP technology has officially entered the era of pre-trained language models. Language models with the Transformer [Vaswani et al. 2017] structure as the basic framework promote the reasonable allocation of information and computing resources by training a general text representation using a large-scale unlabeled corpus, and then using it for the intelligent processing of downstream text with similar language features. The "pre-training + fine-tuning" or "pre-training + template prompt + fine-tuning" information processing process has officially become the basic paradigm of NLP research in the new era [Liu et al. 2023]. The existing diverse pre-trained models can be roughly divided into autoencoder models, autoregressive models, and sequence-to-sequence models according to their basic architectures. Generally speaking, Autoencoder models only stack Transformer encoder structures, construct loss functions by predicting masked tokens in a sentence and excel at text comprehension tasks. Relevant models include BERT [Devlin et al. 2019], ERNIE [Zhang et al. 2019], and etc. Autoregressive models and sequence-to-sequence models are mainly text generation models.

Autoregressive models merely stack Transformer decoder structures, thus being suitable for unidirectional text continuation. During training, the model needs to predict the probability distribution of the output vocabulary in the vocabulary table based on the input partial vocabulary information. In addition, it has to calculate the loss function by comparing the predicted results with the content of the original sequence, and then complete parameter updating. OpenAI, which was the first to use Transformer decoders for text feature extraction, released the GPT1 model [Radford et al. 2018] in 2018. With excellent generation ability, this model can also be applied to natural language understanding tasks by changing the task paradigm. GPT2 [Radford et al. 2019] inherited the basic structure of GPT1, but improved the input form of training data, making the style of pre-training data and task data more similar. GPT2 utilizes more parameters and larger training texts, and the largest-scale GPT2 model has 1.5 billion parameters, which grant it capabilities of zero shot learning. ChatGPT, which has attracted wide attention from scholars recently, represents an improved version of GPT3 [Brown et al. 2020] as the language infrastructure. GPT3 has 96 Transformer decoders, each containing 180 million trainable parameters. Containing 175 billion parameters, the GPT3 model uses 45TB of data during training. As one of the largest language models, GPT3 can model all NLP tasks generatively. Even in the case of providing small samples or no sample at all, the GPT3's performance in answering questions is close to or exceeds the upper limit of the fine-tuning ability of small-to-medium-sized models.

As another major type of generative model, sequence-to-sequence models share the Transformer structure as a whole or use separate encoder and decoder structures to extract text features. With the inclusion of an encoder structure, the language understanding ability of sequence-to-sequence models is generally better than that of common Autoregressive models within a certain range. Representative achievements in this type of model include the MASS [Song et al. 2019] and Unilm [L. Dong et al. 2019] models released by Microsoft, the BART [Lewis et al., 2019] model proposed by Facebook, and the T5 [Raffel et al. 2020] model proposed by Google. Sequence-to-sequence models are also commonly selected architectures in large model production. Related studies such as the FLAN-T5 [Chung et al. 2022] and T0 [Sanh et al. 2022] models have re-optimized the pre-training of the T5 model. One of their features includes using instruction tuning to achieve contextual learning with small sample data.

2.2 Applications of Generative Pre-trained Language Models in Intelligent Processing of Ancient Texts

Generative pre-trained models can be combined with ancient text corpora or multilingual parallel corpora to perform automated translation, text summarization, automatic question answering, text completion, and other generation tasks. They can also be used to perform natural language understanding tasks such as text classification, information extraction, and text retrieval by restructuring the input and output patterns of these tasks. The following sections will introduce the relevant research on the applications of generative pre-trained models in processing ancient texts.

2.2.1 Generative Models and Ancient Text Generation

Generating ancient texts is one of the core applications of generative models. Ancient text generation tasks can be classified into single-language generation and cross-language generation depending on the generation target. Single-language generation requires that the generated text is in the same category as the original text, while cross-language text generation is the opposite. In fact, due to the significant differences between the ancient and modern forms of many languages, intralingual text translation can be viewed as a special case of cross-language text generation.

Examples of single-language generation tasks include generating poetry and ancient language. Relevant studies include [Liao et al. 2019] who used GPT models to generate Chinese classical poetry in different styles. [Hu & Sun. 2020] built a unified framework for generating Chinese classical poetry based on the GPT2 model, using a form of weighted emphasis to control the style of the generated text. The relevant results were included in the "Jiuge" [Guo et al. 2019] poetry generation system developed by Tsinghua University. [Nguyen et al. 2021] developed a Vietnamese poetry generation model called SPGPT2 based on GPT2, which generates Vietnamese traditional poems that conform to the Luc Bat format by adding constraints.

Cross-language text generation tasks include ancient text translation and text summarization. [Yang et al. 2021] and [Tian et al. 2021] exploited the UNILM framework to load GUWEN-BERT [Ethan, 2020/2023] and AnchiBERT [Xujiacheng127/Anchi-Bert · Hugging Face, n.d.] models pre-trained on simplified ancient Chinese data for cross-language text generation tasks. Comparative experimental results showed that adding pre-training mechanisms can effectively improve the model's generation ability. [Chang et al. 2021] designed a translation interface that is sensitive to temporal information based on the GPT2 model to address translation quality problems caused by temporal differences in ancient texts. [Jin et al. 2022] constructed a bilingual parallel corpus of Chinese classical poetry and modern Chinese translations. By use of neural machine translation and trained Transformer and GPT2 models based on the translation data, they achieved automated translation of Chinese classical poetry. [Peng et al. 2021] constructed a historical text dataset in German and Chinese and proposed an algorithm for summarizing ancient historical texts into modern written language using cross-language transfer technology.

2.2.2 Generative models and text completion

Owning to their pre-training tasks, generative pre-training models, including language models and multimodal models, can repair noisy targets and complete missing information. Some researchers utilize generative models to restore the ancient texts which suffered physical damages for various reasons. For example, [Fetaya et al. 2020] trained a recursive neural network with digitized cuneiform corpora from Mesopotamia to repair damaged Babylonian texts. [Assael et al. 2022] used Transformer as the basic structure to develop a model to restore damaged ancient Greek inscriptions. The model accepted text at the level of single characters and designed other structures to model the temporal and geographical characteristics of the text, achieving restoration results better than human experts. [Zheng et al. 2023] developed a double-branch character restoration network based on generative adversarial networks. They exploited image data to train two branches to extract the basic features of the damaged characters and example characters, achieving good restoration results.

In addition to generating or complete ancient texts, generative pre-training models may even change the basic paradigm of natural language understanding tasks. The current label classification problem may gradually shift to specific character generation problems. In theory, a large enough generative model can be applied to all natural language processing tasks, but this requires the joint efforts of multidisciplinary researchers to share discipline data and task descriptions. However, the use of Autoencoder models, represented by BERT, for knowledge extraction in ancient texts remains the mainstream in datamining studies of ancient books. More attention deserves to be paid to the application of generative models in research related to the mining of ancient texts.

3 RESEARCH METHODS

In this chapter, we focus on the key issue of enhancing the ability of GPT-like models to process ancient texts, and introducing the data used in the pre-training research, the pre-training methods, and the methods used to validate the experimental results.

3.1 Pre-training Data Source

The data used for pre-training is the Siku Quanshu of Wenyuan Ge version, which is a large series of books compiled during the reign of Emperor Qianlong in the Qing Dynasty, including four parts: "Jing" Part (经部, classical literature), "Shi" part (史部, historical literature), "Zi" part (子部, ideological literature), and "Ji" part (集部, literary works). Previously, our research team constructed the SikuBERT and SikuRoBERTa models based on the Wenyuan Ge Siku Quanshu, which showed superior performance in tasks such as ancient Chinese part-of-speech tagging, text segmentation, and named entity recognition [Wang et al. 2022]. In this study, we will also investigate whether using the text of the Siku Quanshu to pretrain a generative model can improve the processing performance of text generation tasks in ancient text processing.

As the largest collection of ancient texts in China, the Siku Quanshu consists of 3,461 works from the pre-Qin period to the early Qing Dynasty. It is not only a vital resource for studying ancient Chinese classics, but also an important reference for understanding Chinese culture and history. In this study, we choose the traditional Chinese format of the Siku Quanshu texts as the training corpus to better align with the original texts of ancient books. The total number of characters in the collection is nearly 800 million. We removed the annotation information from the digitized Siku Quanshu text and retained the remaining 53,609,758 Chinese characters for training the GPT2 model, in order to ensure the model's generalization ability on a wider range of ancient Chinese text types.

3.2 Pre-training Method

3.2.1 Pre-training Model Selection

In this study, we chose the GPT2 model as our foundational model for pre-training. GPT2, an autoregressive model designed for natural language generation, is an open-source model released by OpenAI. It demonstrates impressive performance among models with a similar parameter scale. We did not use models with larger parameters, primarily considering deployment issues in practical application scenarios. A GPT2 model specifically fine-tuned with targeted data can achieve effects comparable to larger models, while also benefiting from lower memory usage and faster response times. In the era of large language models, GPT2 remains a highly cost-effective solution. The pre-training models selected in this experiment are as follows: (1) GPT2-Chinese-ancient model [Uer/Gpt2-Chinese-Ancient · Hugging Face, n.d.] open-sourced by Huggingface, which was trained on ancient Chinese literature from the Daizhige library, with a total of 3,000,000 Chinese characters. The training process is based on the UER [Zhao et al. 2019] open-source framework, and the model consists of 12 layers with a vocabulary size of 25,370. (2) GPT2-base-Chinese [Ckiplab/Gpt2-Base-Chinese· Hugging Face, n.d.] model introduced by the CKIP Lab, which is a GPT2-like model designed for traditional Chinese.

3.2.2 Pre-training method selection

As a typical autoregressive language model, GPT has a unidirectional Transformer decoder structure and adopts the training method of the causal language model (CLM). The causal language model is a pre-training task for training unidirectional text representations, in which the model only needs to predict the next word based on the vocabulary on one side of the input sentence and then use a cross-entropy loss function to update the model parameters. Fig 1. illustrates the basic principle of the CLM pre-training method.

As is shown in Fig 1., for an input ancient Chinese sentence, a tokenizer is first used to serialize the sentence, and the serialized result is sent to the embedding layer and multiple Transformer Decoder layers of the GPT2 model for text encoding. The encoded vectors are then used to predict the probability of the next token. The cross-entropy loss function measures the difference between the model's predicted probability distribution and the true distribution. When calculating the loss function, GPT2 typically uses a masking mechanism,

considering only the current token sequence to calculate the maximum probability of the next token, which ensures that the model can only use previous information when generating text. Finally, the backpropagation algorithm is used to calculate the gradient of the loss function with respect to the model's parameters, which are then updated.

Compared with masked language model, the causal language model only allows reference to one side of the content for prediction. When the training goal is to learn a good representation of the input text, masked language model (MLM) is undoubtedly a better choice due to its ability to consider context simultaneously. However, when the training goal is to generate fluent text, the unidirectional causal language model is similar to human writing and can better improve the model's creativity. In this experiment, the training task adopts the causal language model (CLM) and is completed by using the Transformers framework provided by Huggingface company.

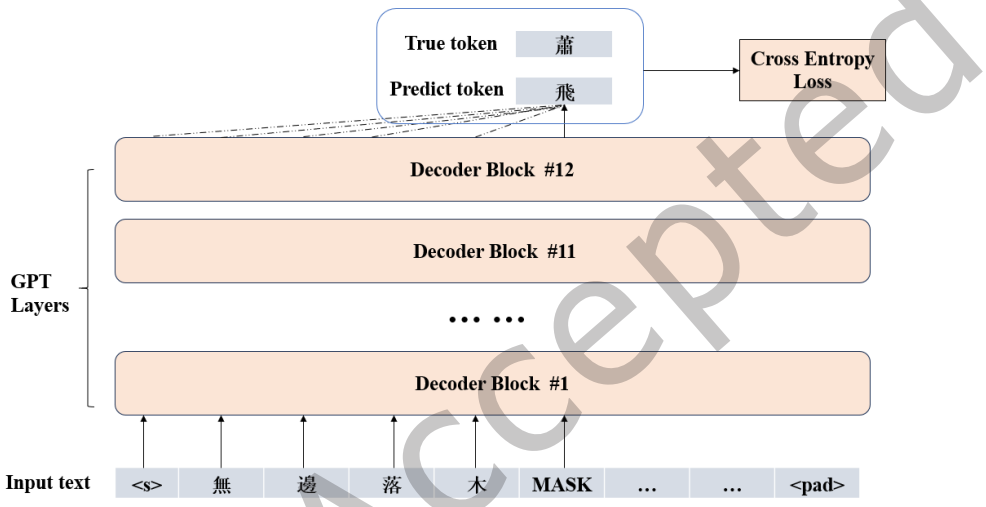


Figure 1 The basic principle of the CLM pre-training method

3.3 Downstream task design

To verify the performance of the SikuGPT pre-training model, this study selects two GPT models designed for ancient Chinese text processing, GPT2-Chinese-ancient and GPT2-base-Chinese. They are chosen as baseline models to further verify their performance in two natural language processing tasks: ancient text translation and ancient text classification.

In downstream task validation, two types of manually processed traditional Chinese classical literature data are used to fine-tune the model on the same computing device, and general computational metrics are employed to evaluate the processing capabilities of different models for downstream tasks. Additionally, the study compares the translation preferences of generative models of different sizes and the affinity of several generative models for evaluation towards different prompts, in order to illustrate the applicability of the SikuGPT model and other generative models.

4 EXPERIMENTAL PROCEDURE AND RESULTS

4.1 Pre-training Experiment

All the Chinese characters that can be displayed in utf-8 encoding were extracted from the entire text of the Siku Quanshu, with 5086 characters added to the GPT2-chinese-ancient model's vocabulary. The filtered Siku Quanshu text was divided into a training set and a validation set in a ratio of 99:1. The model was then fine-tuned using the CLM method based on the Transformers framework. The model training parameters are shown in Table 1.

Table 1 Key hyperparameters for model training

Hyperparameters	Value
learning_rate	5e-5
num_train_epochs	3
per_device_train_bach_size	8
max_seq_length	1024

For preliminary evaluation of the model's performance, perplexity was chosen as the evaluation metric. Perplexity is a language performance evaluation metric based on the probability of sentences in the validation set. The rationale is that a language model performs better if it assigns higher probability values to sentences in the validation set. Since the sentences in the validation set are all normal sentences, a well-trained language model that assigns higher probabilities to the sentences in the validation set indicates a better fitting ability of the model. In perplexity calculation, a sentence can be represented as:

$$S = W_1, W_2, W_3 \dots \dots, W_n \tag{1}$$

The appearance probability of a sentence is:

$$P(S) = P(W_1, W_2, \dots, W_n) = P(W_1)P(W_2|W_1) \dots \dots P(W_{n-1}|W_1, W_2, W_3 \dots, W_n) \tag{2}$$

The formula for calculating perplexity is based on the formula for calculating the probability of a sentence:

$$PPL = \sqrt[n]{\frac{1}{P(W_1, W_2, \dots, W_n)}} \tag{3}$$

We used three GPUs of type RTX 8000 to complete the pre-training task of the model, the whole process took 3 days. The perplexity score of the pre-trained SikuGPT model is 20.85, which does not appear to be a good indicator of performance. However, compared to Autoencoder models, Autoregressive models typically have higher perplexity scores. Moreover, perplexity is not the only standard for evaluating model performance, but rather merely one of the initial indicators, and the performance of the model needs to be explained through its performance in downstream tasks.

4.2 Downstream Task Verification

4.2.1 Text Translation Task

Machine translation is an important task in the field of natural language processing. The language model's ability to understand language can indirectly reflect the model's machine translation performance. At the same time, machine translation can test the model's generation ability. There is still much room for improvement in machine translation for ancient texts because of their unique grammar and vocabulary, especially traditional Chinese texts. As an essential part of Chinese classic literature, standard language and outstanding literary attainments make Twenty-Four Histories an excellent corpus for machine learning testing. This study selects the ancient Chinese text alignment corpus of Twenty-Four Histories as the evaluation corpus. BLEU is adopted as the evaluation metric to test the model's understanding and generation ability.

Based on the CLM task, the pre-training language model SikuGPT can predict the probability distribution of the next word or sequence given the previous context information. Similar to language modeling, the goal of the CLM task is to learn contextual and semantic information of language, so the pre-trained model trained on the CLM task usually has good language understanding and generation abilities. To test the practical effectiveness of the SikuGPT pre-training model, this study used the twenty-four histories parallel corpus of ancient and modern Chinese as experimental data. Afterwards, the machine translation task was exploited to evaluate the model's actual performance. Experimental controls included the GPT2-chinese-ancient and GPT2-base-chinese pre-trained models, as well as the traditional single-layer Transformer model.

(1) Data and Task Description

The data in this study consists of ancient Chinese sentences and their corresponding modern Chinese translations from the Twenty-Four Histories. The Twenty-Four Histories is the general term for the twenty-four official Chinese historical records compiled by scholars in various dynasties in ancient China. It covers nearly 5,000 years of history from the Yellow Emperor period to the Ming Dynasty, and contains a diversity of ancient Chinese culture, including politics, economy, military, and thoughts and etc., making it a valuable

cultural heritage of human civilization. The aligned corpus in the text translation task in this study comes from China’s “The 11th Five Year Plan” key work, “The Complete Translation of the Twenty-Four Histories(《二十四史全译》)”. This was jointly compiled by more than 200 experts in ancient books research over a period of 13 years, representing the highest quality of modern Chinese translations of the Twenty-Four Histories. OCR (Optical Character Recognition) technology and the Aligner alignment software was utilized to align the ancient Chinese and modern Chinese texts at the sentence and paragraph levels. Subsequently, the aligned sentences were filtered based on named entity recognition technology and text matching technology to ensure accurate entity annotation in both ancient and modern Chinese sentences. The aligned sentences with a similarity score between ancient and modern Chinese sentences of 0.85 to 0.98 were selected according to similarity detection. In total, 300,000 high-quality aligned sentence pairs were selected from nearly 1 million aligned sentence pairs for experimentation.

The character lengths of the ancient and modern Chinese sentences total 9,368,674, and 12,236,739, respectively. Their average sentence lengths reach 30.47 characters, and 39.80 characters, respectively, indicating that ancient Chinese is more concise in description than modern Chinese. In terms of variance, ancient Chinese sentences is 1977.29, while modern Chinese sentences is 3424.18, indicating that there is greater dispersion between sentences in modern Chinese compared to ancient Chinese.

These details are shown in Table 2. The average sentence length of ancient Chinese in the corpus is over 30 characters, which is to enable the model to better learn the relationship between parallel sentence pairs in the corpus. In addition, during the preprocessing of the experimental data, the corpus was divided into a training set and a test set in a 9:1 ratio. Specifically, 270,000 pairs were used as training data, while 30,000 pairs were used as test data. A sample of the aligned ancient-modern Chinese sentence pairs is shown in the Table 2.

Table 2 Basic information of ancient-modern Chinese alignment corpus

	Ancient Chinese	Modern Chinese
Total number of characters	9368674	12236739
Average word count per sentence	30.47	39.80
variance	1977.29	3424.18

Table 3 shows some examples from our corpus that are challenging for both humans and machines. These examples were used in subsequent experiments to finely evaluate the translation performance of different models.

Table 3 Examples of aligned corpus

Ancient Chinese sentence	Corresponding modern Chinese sentence
後與秦戰，爲秦所獲，立十四年而死。	後來與秦國作戰，被秦軍捉住，在位十四年而死。
此有禮於君者，王如堅諸人是也。	這是對待君主有禮的做法，王如堅諸位人士就是這樣的人。
有其人則七，無其人則五。至光武中興，及魏、晋、宋、齊、隋、唐，或立六廟，或立四廟，蓋建國之始，未盈其數也。	有相應的人就建立七座，没有相應的人就建五座。光武中興, 以及魏、晋、宋、齊、隋、唐，有的建立六廟, 有的建立四廟，是因爲建國開始時期，不滿這個數目。

*①The first sentence means “Later, he fought against the state of Qin and was captured by the Qin army. He reigned for 14 years and died.”

②The second sentence means “This is the proper way to treat a monarch, and Wang Rujian is such a person.”

③The third sentence means “If there are corresponding people, seven temples are established; if there are no corresponding people, five are built. During the period of Guangwu’s restoration, as well as the dynasties of Wei, Jin, Song, Qi, Sui, and Tang, some established six temples, while others established four, because at the beginning of the founding of the nation, the number was not enough.”

(2) Model Validation

To validate the performance of SikuGPT, the baseline models selected for the validation experiment are GPT2-chinese-ancient, GPT2-base-chinese, SikuBERT, and Transformer GPT2-chinese-ancient is a pre-trained language model based on GPT2 for generating ancient Chinese text. It is trained on un-punctuated data from the DaiZhige ancient Chinese corpus and then supplemented with punctuation. GPT2-base-chinese, a Chinese version of GPT2 trained on 15GB of Chinese language data, can be exploited for poem generation, news

generation, and novel continuation. SikuBERT is a pre-trained model based on the BERT-base-Chinese architecture and fine-tuned on the Siku Quanshu corpus by Nanjing Agricultural University. With a word list using traditional Chinese without punctuation, sentence segmentation is performed at the character level. To adapt the BERT-based model for text translation tasks, the UNILM framework is adopted to load the model weights and reconstruct the BERT model's input format.

(3) Model Validation Performance Evaluation Metrics

BLEU (Bilingual Evaluation Understudy) is a commonly used evaluation metric in machine translation. BLEU measures the similarity between the machine translation output and the human reference translation to evaluate the quality of machine translation systems. The evaluation result is represented as a score between 0 and 1, with higher scores indicating better machine translation quality. BLEU metric calculation is based on n-gram matching and sentence length penalty. Specifically, the BLEU metric matches the n-grams in the translation output with those in the reference translation and calculates an n-gram matching score based on the number of matches. Meanwhile, the BLEU metric penalizes longer translation results to avoid machine translation systems generating meaningless vocabulary to increase their scores.

(4) Model Hyperparameter Settings

The computer configuration in this experiment is as follows: Operating System: CentOS 3.10.0; CPU: 4 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz; Memory: 256G; GPU: 6 NVIDIA Tesla P40, VRAM: 24G. To ensure the comparability of the validation results, the hyperparameters of the model training are kept consistent in the validation experiment. The key hyperparameters used in text translation tasks are shown in the following table:

Table 4 Key hyperparameters of translation model training

Hyperparameters	Value
max_gen_length	512
max_seq_length	1024
train_batch_size	8
epoches	5
learning_rate	1e-5
warmup_proportion	0.1

(5) Comparison of Text Translation Performance

As shown in Table 5, in terms of Chinese text translation performance, the GPT-based models all surpass the basic Transformer architecture. Both SikuGPT and GPT2-base-chinese models perform better than GPT2-chinese-ancient and SikuBERT+unim models in terms of single-character level and fluency of generated text. Furthermore, the translation performance of the GPT-based models exceeds that of SikuBERT in terms of both single-character level and fluency. Overall, SikuGPT performs the best in the translation task with slight advantages in every BLEU score compared to GPT2-base-chinesegpt2chinese. In the translation task, SikuGPT not only captures the semantic information of words and phrases accurately but also performs better in capturing the semantic information of longer sentences, resulting in more accurate translations.

Table 5 Comparison of translation performance of different models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GPT2-chinese-ancient	0.735	0.546	0.432	0.351
SikuGPT	**0.765**	**0.593**	**0.488**	**0.413**
GPT2-base-chinese	0.761	0.589	0.485	0.410
SikuBERT+UNILM	0.658	0.514	0.423	0.357
Transformer	0.695	0.500	0.379	0.301

To evaluate the specific translation performance of different models, we selected three test sentences from the test set, just as shown in Table 3. Some closed-source large language models were also selected to generate translation text through providing instructions and raw ancient Chinese sentences, including ChatGPT

[ChatGPT, n.d.], ERNIE-3.5 [Wenxinyiyan, n.d.] and Tongyi qianwen2.5[Tongyi, n.d.].The specific translation results are shown in Table 6.

In general, the translation results of most models are relatively close to the original translation. Among them, the SikuGPT's translation quality is superior to other fine-tuned small models and even better than the unmodified ChatGPT3.5. For example, in the translation of sentence 2, ChatGPT3.5 directly provided a translation that deviated from the original semantic meaning, failing to correctly understand "王如堅" as a person name, but rather interpreting it as "堅強的王者（a strong king）." Compared to large language models, SikuGPT's translation results focus on the accurate expression of the original content, which is reflected in the model's reluctance to add omitted content to the translation. However, the four large language models have almost all supplemented the omitted components or pronouns in the original text, although sometimes these supplements are not correct. For instance, when translating the third sentence, Tongyi qianwen2.5 directly translated "其人" as "君王（monarch）," and all the large language models understood "其数" as a specified number. This supplementary behavior makes the translation results of large language models more readable and coherent, suitable for those whose mother tongue is modern Chinese. We speculate that this difference may be determined by the amount of modern Chinese corpora used during the pre-training phase, with models trained on a large volume of modern Chinese text tending to imitate the grammar and expression habits of modern Chinese when performing translation tasks.

Table 6 Translation results

Model	Translation results
GPT2-chinese-ancient	(1) 後來和秦軍作戰，被秦軍俘獲，立下十四年死去。 (2) 這是對於君主的人，王如堅等人就是。 (3) 有這個人就是七廟，没有那些人就是五廟。到光武年間,到魏、晋、宋、齊、隋、唐，或立六廟，或建四廟，大概是建國的開始，不超過數量。
SikuGPT	(1) 後來與秦軍交戰，被秦國擒獲，在位十四年死去。 (2) 這是對君主有禮的，王如堅等人就是。 (3) 有這個人就是七廟，没有那些人就是五廟。到光武中興,及魏、晋、宋、齊、隋、唐時期,或立六廟，或建四廟，大概是創建國家的開始，還不超過其數
GPT2-base-chinese	(1) 後來和秦作戰，被秦俘獲，在位十四年而死。 (2) 這是對於君主有禮的，王如堅决衆人。 (3) 有這個人就是七個，没有其人便是五個。到光武年間興建,以及魏、晋、宋、齊、隋、唐，或立六廟，或者建立四廟，大概是建國之始，不滿足數字的原因。
ChatGPT3.5(zero-shot)	(1) 後與秦國交戰，被秦國俘虜，後來在秦國度過了十四年並最終去世。 (2) 對君主表示禮敬的人，就像對待堅強的王者一樣。 (3) 有人就有七廟，無人則為五廟。從光武中興到魏晉宋齊隋唐，或建立六廟，或建立四廟，這是建立國家的開始，數目尚未完全。
ChatGPT4(zero-shot)	(1) 在與秦國作戰之後，被秦國俘虜，當了十四年國君，最後去世。 (2) 對君主有禮的人，是王如堅等人。 (3) 如果有合適的人選，則設置七廟；如果沒有合適的人選，則設置五廟。從光武帝中興之後，經過魏、晉、宋、齊、隋、唐等朝代，有時立六廟，有時立四廟，這是因為建國之初，尚未達到規定的數目。
ERNIE-3.5 (zero-shot)	(1) 後來與秦軍交戰，被秦軍俘獲，在位十四年而死。

	(2) 這就是對君王有禮的人，像王如堅這類人就是這樣。 (3) 有其人則設立七廟，無其人則設立五廟。到了光武帝中興之時，以及魏、晉、宋、齊、隋、唐這些朝代，有的設立六廟，有的設立四廟，大概是在建國之初，尚未達到常規的數量。
Tongyi qianwen2.5(zero-shot)	(1) 後來與秦國交戰，被秦國俘獲，他在位十四年就去世了。 (2) 這就是對君主有禮貌的人，像王堅那樣的人就是例子。 (3) 若君王在位，則宗庙祭祀有七代先祖；若君王不在，則縮減至五代。到了光武帝中興漢室，以及後續的魏、晉、宋、齊、隋、唐等朝代，有的設立六座宗庙，有的僅設四座，這都是因為在這些國家建立之初，尚未能達到祭祀七代祖先的完整規制。

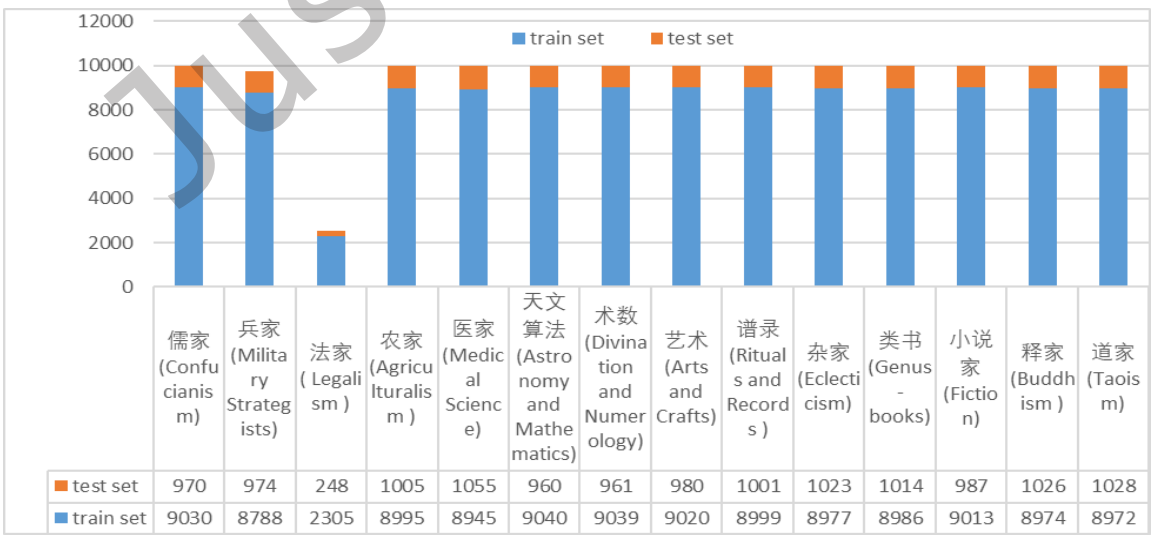
4.2.2Text Classification Task

As a basic task in the field of natural language processing, text classification aims to classify a given text into predefined categories. Unlike machine translation tasks, text classification tasks are simpler and more direct. On the one hand, text classification tasks usually require understanding the semantics of the text. Text classification tasks can be applied to testing the model's ability to understand natural language meanings. On the other hand, the model proposed in this study is a language model obtained through unsupervised training.

To further validate the performance of SikuGPT pre-training model, a text classification task was also conducted in this study as a downstream validation experiment. In this experiment, the required training and testing data was constructed based on the 14 categories of ancient texts in the sub-parts of Siku Quanshu. Several pre-trained models were introduced as baseline models for the validation experiment, including GPT2-chinese-ancient and GPT2-base-chinese as well as SikuRoberta, which was specifically fine-tuned for the ancient text classification task.

(1) Corpus and Task Description

Siku Quanshu classified ancient Chinese books into four categories, and the “Zi” part were further divided into 14 sub-categories, including Confucianism, Military Strategists, Legalism, Agriculturalism, Medicine, Astronomy and Mathematics, Arts, Divination and Numerology, Rituals and Records, Miscellaneous Works, Novels, Buddhism, Taoism, and Genus-books. The corpus used in the text classification task was the traditional Chinese version of Siku Quanshu data collected through web crawlers [Hu et al. 2022]. After preliminary data preprocessing, a total of 132,315 valid text data containing all 14 categories were obtained. Afterwards, the data was divided into training and testing sets in a ratio of 9:1, with 119,083 as training data and 13,232 as testing data. The distribution of training and testing data in each sub-category is shown in the following figure:



	儒家 (Confucianism)	兵家 (Military Strategists)	法家 (Legalism)	农家 (Agriculturalism)	医家 (Medical Science)	天文算法 (Astronomy and Mathematics)	术数 (Divination and Numerology)	艺术 (Arts and Crafts)	谱录 (Rituals and Records)	杂家 (Eclecticism)	类书 (Genus-books)	小说家 (Fiction)	释家 (Buddhism)	道家 (Taoism)
test set	970	974	248	1005	1055	960	961	980	1001	1023	1014	987	1026	1028
train set	9030	8788	2305	8995	8945	9040	9039	9020	8999	8977	8986	9013	8974	8972

Figure 2 Distribution of text classification task data set across sub-categories

The text classification task aims to make the GPT pre-training model automatically generate the category labels correctly through training fine-tuning on the above corpus. Generally, the idea of utilizing pre-trained models for text classification is to input the output vector of the model's last layer into the fully connected layer to calculate the probability of each category. For GPT pre-training models, the output vector of the model is the vector corresponding to the last token of the input sequence. From the training task, this vector represents the semantic information of the token which is most likely to correspond to the next character under the condition that the first n tokens are determined. If directly exploited for text classification, this vector may cause a problem of inconsistency between the downstream task and the pre-training task structure. Therefore, the research team borrowed the idea of prompt learning and used the generative modeling approach to restructure the task paradigm of text classification. Specifically, two concatenation prompt sentences were separately added to the input text to guide the model to decode the text of the label itself, thus converting a classification task into a generation task, and ensuring the consistency of the upstream and downstream training tasks. Template design method are shown in Table 7:

Table 7 Prompt template design

Original statement	Input statement
[CLS]text[SEP]	[CLS]text[SEP] 这段文本的类别是____
[CLS]text[SEP]	[CLS]给定标签列表[…]，请根据列表中的元素为以下文本选择合适的分类标签: text这段文本的类别是____

*"这段文本的类别是____"means " The category of this text is___".
"给定标签列表[…]，请根据列表中的元素为以下文本选择合适的分类标签: "means " Given the list of labels […], please select an appropriate category label for the following text based on the elements in the list:".

(2) Model performance evaluation index

For the performance evaluation of each category in text classification, precision (P), recall (R), and F1-score are adopted as evaluation metrics. Weighted precision (Weighted_P), weighted recall (Weighted_R), and weighted F1-score (Weighted_F) are used to calculate the overall classification performance.

(3) Comparison of text classification effect

According to Table 8, in terms of overall classification performance, SikuGPT using the first prompt template is the best, with an F1 score reaching 90.50, surpassing other generative models., and achieved a slight advantage over the 1.0 version of the SikuRoberta model trained with fine-tuning. This phenomenon indicates that changes in the prompt templates have a significant impact on the classification performance of the models. One possible explanation is that different prompt templates affect the way in which different small models utilize internal knowledge. Prompt template 1 may be more aligned with the pre-training objectives of SikuGPT and GPT2-chinese-ancient, while prompt template 2 may be more in line with the pre-training objectives of GPT2-base-chinese. Overall, since SikuGPT has been trained on a larger amount of ancient literature data, its understanding of modern text may be relatively weakened. When using prompt templates, it is important to control the length of modern Chinese text to more easily achieve good classification results.

Table 8 Overall classification effect comparison of each model

Model	Precision（%）	Recall（%）	F1-score（%）
GPT2-chinese-ancient(prompt1)	89.63	89.65	89.59
GPT2-base-chinese(prompt1)	90.30	90.27	90.26
SikuGPT(prompt1)	**90.51**	**90.54**	**90.50**
GPT2-chinese-ancient(prompt2)	89.52	89.47	89.44
GPT2-base-chinese(prompt2)	90.49	90.49	90.47
SikuGPT(prompt2)	90.39	90.40	90.36
SikuRoberta _v1(non-prompt)	86.37	94.60	90.21

The detailed precision, recall, and F1 scores of SikuGPT using the first prompt template on each category are shown in Table 9. Among the 14 categories, SikuGPT obtained high scores in categories such as astronomy, medicine, and agriculture, while performing relatively poorly in categories such as miscellaneous literature, Taoism, and Legalism. This may be related to the characteristics of the texts in each category. The texts in the astronomy category often exhibit salient features in describing astronomical phenomena, such as “平帝元始元年辛酉歲五月丁巳朔日食” (This sentence means: “The solar eclipse occurred in the early morning of Dingsi day in May of Xinyou year in the first year of Yuanshi of Emperor Ping.”). By contrast, the texts in the miscellaneous literature and Taoism categories encompass a wide range of content and structure, without exhibiting obvious features in text content and structure. Additionally, the poor performance of the Legalism category may be due to the small size of the training set, which affected the model's performance on the classification task.

Table 9 Classification effect of each category of the best performing model SikuGPT

Category	Precision	Recall	F1-score
儒家（Confucianism）	83.69	89.38	86.44
兵家（Military Strategists）	91.99	89.63	90.80
农家（Agriculturalism）	93.22	94.33	93.77
医家（Medical Science）	93.42	96.87	95.11
天文算法（Astronomy and Mathematics）	97.78	96.46	97.12
小说家（Fiction）	87.98	86.02	86.99
术数（Divination and Numerology）	90.36	94.59	92.43
杂家（Eclecticism）	79.20	73.31	76.14
法家（Legalism）	85.49	87.90	86.68
类书（Genus-books）	94.58	91.22	92.87
艺术（Arts and Crafts）	93.72	94.49	94.11
谱录（Rituals and Records）	92.08	91.81	91.95
道家（Taoism）	86.06	86.48	86.27
释家（Buddhism）	94.12	93.57	93.84

5 DISCUSSION

The validation experiments in this study demonstrate the relative superiority of SikuGPT in downstream tasks such as text translation and text classification, providing the following insights:

5.1 Model domain specialization: an effective strategy for vertical domain tasks

In this study, pre-training GPT2-chinese-ancient was continued on the ancient Chinese corpus of the Siku Quanshu to obtain SikuGPT, which showed significantly improved performance in natural language processing tasks such as ancient Chinese text translation and text classification compared to the baseline model. This result indicates that using domain-specific corpora to continue pre-training generative models can effectively improve their performance in executing downstream tasks on texts with similar linguistic and domain characteristics. Furthermore, in this study the original vocabulary of GPT2-chinese-ancient was enlarged with over 5000 traditional Chinese characters. The purpose was to enhance the model's encoding ability for ancient Chinese texts and avoid the model's inability to recognize rare characters and generate invalid characters such as “[UNK]”. Currently, improvement of models’ performance heavily relies on the quality and scale of training corpora. The superior performance of SikuGPT in this study further demonstrates that building domain-specific models represents an effective and necessary means of improving model performance on vertical domain tasks. On the basis of the huge success achieved by universal generative language models, large models specifically designed for a given domain will show even greater performance advantages. However, the construction of high-quality large-scale corpora at lower costs still poses an essential research problem in current information resource management disciplines.

Through comparative research, we have found that in the processing of Chinese ancient books, large language models like ChatGPT3.5, which are fine-tuned with general instruction data, often fail to effectively address issues. While models with ultra-large parameters, such as GPT4, can achieve desirable results in handling target texts, they incur higher usage costs and often face issues with connection stability. In this context, to obtain quick and high-quality responses, using a smaller, domain-specific model fine-tuned with specialized corpora presents a more cost-effective solution for scholars with limited computing resources.

5.2 Generative pre-training models: a new paradigm for intelligent processing of classics

In this study, a generative model was applied to text translation, which promoted the innovation of the research paradigm for text translation. Traditional translation models consist of two core modules: an encoder that encodes the source language text to a semantic vector and a decoder that receives the vector and decodes it into natural language text expressed in the target language. Models based on the encoder-decoder architecture require a large amount of parallel corpora consisting of source and target language texts during the training phase. Consequently, the effect of the translation model depends heavily on the granularity and scale of the aligned parallel corpus. In contrast, the various GPT models used in this study serve as decoders with a simpler structure than those of traditional translation models. Moreover, the construction of the translation model in this study was realized through two steps of continued pre-training and fine-tuning. During pre-training, only the traditional Chinese corpus of Siku Quanshu was utilized, avoiding the cumbersome task of constructing a large number of parallel corpora while improving the model's performance on translation tasks. It is foreseeable that if bilingual or multilingual corpora are exploited during pre-training, the model's performance on downstream tasks such as translation can be further improved.

This study also demonstrates the feasibility of applying generative models to text classification tasks. Text classification models have evolved from traditional statistical machine learning models through traditional deep learning models to pre-trained language models such as BERT. The mainstream approach in the past has been to utilize Encoder-based pre-trained models such as BERT combined with an additional output layer to build text classification models. However, models constructed through this approach have limited generality, as they can only be used for text classification and can only classify text into a limited number of categories. In this study, a Decoder-based generative model was adopted to complete the text classification task. The generative model uses the text to be classified as a “prompt" for the model, which then generates content that can serve as a category label. The model achieved accuracy comparable to that of vertical domain models, which demonstrates that generative models can reasonably generate expected content under the joint constraint of the task corpus and prompt template. Additionally, for the original pre-trained language model, this model can effectively avoid its semantic representation ability being weakened during the fine-tuning stage in the traditional “pre-training + fine-tuning” mode. This study also demonstrates that generative models can be employed to complete text classification tasks without the need for additional output layers. This stands in contrast to the mainstream approach using pre-trained Encoder-based models such as BERT. Generative models can be applied to multiple tasks, including text translation and text classification, and have a wide range of applicability. This capability may reshape the paradigm of organizing and serving knowledge of ancient texts.

6 CONCLUSION

In this study, the SikuGPT pre-training model was developed by incorporating the corpus of Siku Quanshu into the GPT model. The SikuGPT pre-training model is a practical application of the integration and development of generative pre-training technology and digital humanities research. It not only expands the application field of generative pre-training technology, but also enriches the technical content of digital humanities. In light of this, the study has major theoretical and practical significance for ancient text processing and digital humanities research.

In the future, we will closely follow the latest development in generative AI research, and explore the training of pre-training models with larger parameter scales and efficient parameter fine-tuning. In addition, new technologies and concepts need to be actively integrated in the field of natural language processing to deeply explore the value and knowledge of traditional Chinese culture. Furthermore, a knowledge service

platform can be developed based on large pre-training models for ancient Chinese. With such platform, multiple NLP tasks of ancient texts can be unified through a question-answering system, thus providing users with high-quality ancient knowledge services. Ultimately, this will promote the dissemination of traditional Chinese culture and the knowledge contained in ancient Chinese texts. Finally, we open source the SikuGPT model to github (https://github.com/SIKU-BERT/sikuGPT) so that researchers from different professional backgrounds can download it and jointly promote the communication between different cultures.

CONFLICT OF INTEREST STATEMENT

The authors declared that they have no conflicts of interest to this work.

REFERENCES

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, *603*(7900), Article 7900. https://doi.org/10.1038/s41586-022-04448-z

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Chang, E., Shiue, Y.-T., Yeh, H.-S., & Demberg, V. (2021). Time-Aware Ancient Chinese Text Translation and Inference (*arXiv:2107.03179*). arXiv. https://doi.org/10.48550/arXiv.2107.03179

ChatGPT. (n.d.). Retrieved May 17, 2024, from https://openai.com/chatgpt/

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2022). Scaling Instruction-Finetuned Language Models (*arXiv:2210.11416*). arXiv. https://doi.org/10.48550/arXiv.2210.11416

Ckiplab/gpt2-base-chinese · Hugging Face. (n.d.). Retrieved April 8, 2023, from https://huggingface.co/ckiplab/gpt2-base-chinese

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (*arXiv:1810.04805*). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified Language Model Pre-training for Natural Language Understanding and Generation. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html

Ethan. (2023). Ethan-yt/guwenbert. https://github.com/Ethan-yt/guwenbert (Original work published 2020)

Fetaya, E., Lifshitz, Y., Aaron, E., & Gordin, S. (2020). Restoration of fragmentary Babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, *117*(37), 22743–22751. https://doi.org/10.1073/pnas.2003794117

General Catalog of Chinese Ancient Books. (n.d.). Retrieved May 13, 2024, from https://bibliographical.guji.cn/

Hu, J., & Sun, M. (2020). Generating Major Types of Chinese Classical Poetry in a Uniformed Framework (*arXiv:2003.11528*). arXiv. https://doi.org/10.48550/arXiv.2003.11528

Hu H, Zhang Y, Deng S, Wang D, Feng M, Liu L, & Li B. (2022). Automatic Text Classification of “Zi” Part of Siku Quanshu from the Perspective of Digital Humanities：Based on SikuBERT and SikuRoBERTa Pre-trained Models. *LIBRARY TRIBUNE*, *42*(12), 138–148.

Jin, P., Wang, H., Ma, L., Wang, B., & Zhu, S. (2022). Translating Classical Chinese Poetry into Modern Chinese with Transformer. In M. Dong, Y. Gu, & J.-F. Hong (Eds.), *Chinese Lexical Semantics* (pp. 474–480). Springer International Publishing. https://doi.org/10.1007/978-3-031-06703-7_37

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (*arXiv:1910.13461*). arXiv. https://doi.org/10.48550/arXiv.1910.13461

Liao, Y., Wang, Y., Liu, Q., & Jiang, X. (2019). GPT-based Generation for Classical Chinese Poetry (*arXiv:1907.00151*). arXiv. https://doi.org/10.48550/arXiv.1907.00151

Lin L., & Wang D. (2023). A Survey of Ancient Book Text Mining Technology. Scientific Information Research, 5(1), 78–91. https://doi.org/10.19809/j.cnki.kjqbyj.2023.01.006

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 195:1-195:35. https://doi.org/10.1145/3560815

Moradi, M., Blagec, K., Haberl, F., & Samwald, M. (2022). GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain (*arXiv:2109.02555*). arXiv. https://doi.org/10.48550/arXiv.2109.02555

Nguyen, T., Nguyen, P., Pham, H., Bui, T., Nguyen, T., & Luong, D. (2021). SP-GPT2: Semantics Improvement in Vietnamese Poetry Generation. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1576–1581. https://doi.org/10.1109/ICMLA52953.2021.00252

Peng, X., Zheng, Y., Lin, C., & Siddharthan, A. (2021). Summarising Historical Text in Modern Languages. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*: Main Volume, 3123–3142. https://doi.org/10.18653/v1/2021.eacl-main.273

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 140:5485-140:5551.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., … Rush, A. M. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization (*arXiv:2110.08207*). arXiv. https://doi.org/10.48550/arXiv.2110.08207

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2019). MASS: Masked Sequence to Sequence Pre-training for Language Generation (*arXiv:1905.02450*). arXiv. https://doi.org/10.48550/arXiv.1905.02450

Sturgeon, D. (n.d.). Chinese Text Project. Retrieved May 13, 2024, from https://ctext.org/

Tian, H., Yang, K., Liu, D., & Lv, J. (2021). AnchiBERT: A Pre-Trained Model for Ancient Chinese Language Understanding and Generation. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534342

Tongyi. (n.d.). Retrieved May 17, 2024, from https://tongyi.aliyun.com/qianwen/

Uer/gpt2-chinese-ancient · Hugging Face. (n.d.). Retrieved April 8, 2023, from https://huggingface.co/uer/gpt2-chinese-ancient

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang D, Liu C, Zhu Z, Liu J, Hu H, Si S, & Li B. (2022). Construction and Application of Pre-trained Models of Siku Quanshu in Orientation to Digital Humanities. *LIBRARY TRIBUNE*, *42*(6), 31–43.

Wenjun, Z., Benpeng, S., Ruiqi, F., Xihua, P., & Shanxiong, C. (2023). EA-GAN: Restoration of text in ancient Chinese books based on an example attention generative adversarial network. *Heritage Science, 11*(1), 42. https://doi.org/10.1186/s40494-023-00882-y

Wenxinyiyan. (n.d.). Retrieved May 17, 2024, from https://yiyan.baidu.com/welcome

White Paper on Artificial Intelligence-Generated Content (AIGC) (2022) -- China Information and Communication Academy. Retrieved March 28, 2023, from http://www.caict.ac.cn/sytj/202209/t20220913_408835.htm

Xujiacheng127/anchi-bert · Hugging Face. (n.d.). Retrieved April 8, 2023, from https://huggingface.co/xujiacheng127/anchi-bert

Yang, Z., Chen, K., & Chen, J. (2021). Guwen-UNILM: Machine Translation Between Ancient and Modern Chinese Based on Pre-Trained Models. In L. Wang, Y. Feng, Y. Hong, & R. He (Eds.), *Natural Language Processing and Chinese Computing* (pp. 116–128). Springer International Publishing. https://doi.org/10.1007/978-3-030-88480-2_10

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities (*arXiv:1905.07129*). arXiv. https://doi.org/10.48550/arXiv.1905.07129

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Chen, X., Deng, H., Ju, Q., & Du, X. (2019). UER: An Open-Source Toolkit for Pre-training Models (*arXiv:1909.05658*). arXiv. https://doi.org/10.48550/arXiv.1909.05658

Zhipeng, G., Yi, X., Sun, M., Li, W., Yang, C., Liang, J., Chen, H., Zhang, Y., & Li, R. (2019). Jiuge: A Human-Machine Collaborative Chinese Classical Poetry Generation System. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 25–30. https://doi.org/10.18653/v1/P19-3005