npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01747-2

HsscBERT: pre-training domain model for the full text of Chinese humanity and social science

Check for updates

Si Shen[1] ✉, Zhixiao Zhao[2], Dayu Yan[1], Chuan Jiang[2] & Dongbo Wang[2]

Advancements in natural language processing (NLP) have driven research in content evaluation, but no pre-trained models exist for Chinese humanities and social sciences. To address this, several BERT-based pre-trained models for these fields were trained and validated. The results show improved performance in tasks like text classification, structural function recognition, and named entity recognition. These models enhance the intelligent processing of Chinese humanities and social science texts, supporting cross-lingual model development.

Natural language processing (NLP) techniques are essential for text processing, with advancements in deep learning enabling robust results in tasks like word segmentation, entity recognition, and text classification[1–3]. Pre-training models have revolutionized NLP, originating from the introduction of word embeddings by Bengio et al.[4] in 2003. Word2Vec[5] and GloVe[6] improved text processing but struggled with polysemy due to static word vectors. Modern pre-trained models, like GPT[7] (generative) and BERT[8] (understanding), utilize the entire model architecture for downstream tasks. GPT, employing a powerful one-way Transformer[9], excels at capturing long-range dependencies and pioneered unsupervised pre-training followed by supervised fine-tuning[10]. BERT, a deep bidirectional Transformer, leverages Masked Language Modeling (MLM) for contextual understanding and Next Sentence Prediction (NSP) for inter-sentence relationships, significantly advancing the field. A significant amount of research has focused on improving the performance of BERT in various applications by refining pre-training tasks, incorporating external knowledge, and optimizing transformer architectures[11].

The performance optimization studies for the BERT model can be divided into the following aspects. Cross-linguistic knowledge fusion: Hu et al.[12] used multilingual attention for sentiment analysis. Chen et al.[13] fine-tuned on augmented multilingual data for low-resource languages. Caciularu et al.[14] introduced CDLM for multi-document language modeling. Multimodal information fusion: Research emphasizes visual text processing. Xu et al.[15] proposed LayoutLM for layout and content understanding. Luo et al.[16] combined visual and language pre-training models. Li et al. introduced VTR-PTM with dual-stream input for image captioning. Li et al.[17] proposed P-PCFL for precise vision-language connections. Long text processing: Dai et al.[18] introduced Transformer-XL with segment-level recursion. Yang et al.[19] proposed XLNet using an AR model and dual-stream attention. Song et al.[20] developed MPNet by integrating MLM and PLM with location information. Ding et al.[21] presented CogLTX to address limitations in long-range attention. Model distillation and compression: Jiao et al.[22] introduced a transformer-based distillation method. Xu et al.[23] presented Theseus for BERT compression through modular replacement and emulation. He et al.[24] studied KDEP, a feature-based knowledge distillation method.

The integration of specialized domain knowledge has proven crucial for enhancing the generalization ability and controllability of BERT models. This has led to significant research efforts in developing domain-specific pre-trained models, primarily focusing on biomedical and general academic texts. In the biomedical domain, Lee et al.[25] demonstrated the superior performance of BioBERT over standard BERT in tasks like named entity recognition, relationship extraction, and question answering. Jeong et al.[26] further applied BioBERT to biomedical question answering, achieving notable success with a sequential transfer learning approach. BioBERT's effectiveness in extracting medical text features was also highlighted by Xiao et al.[27]. To address fine-grained semantic relationships in this domain, Liu et al.[28] proposed SAPBERT for accurate biomedical entity linking. The urgency of the Covid-19 pandemic spurred the development of CovidBERT by Hebbar and Xie[29] for extracting relevant treatment information, while Rasmy[30] introduced Med-BERT, integrating BERT with structured domain knowledge for improved disease prediction.

Similarly, the academic domain has witnessed the development of specialized BERT models. Beltagy et al.[31] released SciBERT, which outperformed BERT on scientific text, addressing the scarcity of large-scale, high-quality scientific data. Park and Caragea[32] corroborated SciBERT's advantages in intermediate task transfer, scientific keyword recognition, and classification. Van Dongen et al.[33] introduced SChuBERT for citation prediction, leveraging a comprehensive academic corpus with citation links. Liu et al.[34] trained OAG-BERT by incorporating diverse academic entities, demonstrating its effectiveness. Shen et al.[35] developed SsciBERT specifically for English social science abstracts, achieving state-of-the-art results in

[1]Group of Science and Technology Full-text Knowledge Mining, School of Economics & Management, Nanjing University of Science and Technology, Nanjing, China. [2]College of Information Management, Nanjing Agricultural University, Nanjing, China. ✉e-mail: jisuanyuyan@163.com

humanities and social sciences. In the context of Chinese academic literature, Li et al.[36] created a dataset and corresponding models based on CNKI abstracts and titles.

These studies underscore a growing trend in BERT research towards domain-specific enhancements. However, a notable gap exists in pre-trained models tailored for full-text text mining within the humanities and social sciences. While large language models like GPT have broadly advanced NLP, smaller BERT-based models continue to excel in specific domains, exhibiting stronger domain integration in multimodal and cross-lingual applications. Although academic domain-specific models are prevalent in natural sciences, the development for English humanities and social science is recent and primarily focused on abstracts. The creation of a Chinese humanities and social science pre-trained model utilizing full-text academic resources remains an underexplored area.

Methods

In the field of deep learning, the structure of researches using domain data pre-training (e.g., SciBERT and BioBERT) can typically be allocated to two parts—pre-trained experiments and performance validation experiments—conducted on the model using a standard dataset. In addition to inheriting this basic experimental paradigm, to make the content more characteristic of humanity and social science research, this study also made improvements in light of the characteristics of general research within the disciplines of Chinese humanity and social science dataset. The following experimental steps are designed.

(1) Data acquisition and pre-processing. A python crawler was used to crawl the abstracts of papers included in the Chinese Social Sciences Citation database from 1998 to 2020 and the full text of all papers included in the China Social Science Excellence database from 1995 to 2021. Any abnormal characters and spaces in the text were cleaned, the full text of each paragraphs and the abstract as the basic unit of the pre-training data were determined, the abstracts and full academic texts of papers into the same file with line breaks were merged, and the training and validation datasets at a ratio of 99:1 were divided.

(2) Model pre-training. Using the constructed corpus of Chinese humanity and social science abstracts and full-texts, this study trained and developed a pre-trained model specifically designed for Chinese humanity and social science. The training process utilized BERT, which was trained on a general Chinese vocabulary. The pre-trained model was constructed using a self-built GPU server and was named as HsscBERT (https://github.com/S-T-Full-Text-Knowledge-Mining/HsscBERT).

(3) Pre-trained model evaluation. Perplexity was used as an evaluation metric to initially determine the whole performance of the pre-trained models. A standardized performance evaluation experiment is required to ensure the usability and effectiveness of the model. We designed the following model performance validation experiments based on the characteristics of Chinese humanity and social science data. Based on the validation dataset, we conducted text classification, structural function recognition and named entity recognition experiments and selected the benchmark models and the new models in this research for the comparison. Figure 1 presents the basic flow of the experiments.

Pre-train experiment

Pretraining model. Currently, mainstream pre-training models are trending towards larger sizes, more parameters, and increasingly complex structures. In this study, we chose BERT as the baseline model to conduct pre-training using the Masked Language Model (MLM) task.

BERT represents a pioneering advancement in the field of natural language understanding in the deep learning era. What distinguishes BERT from previous pre-training models is its ability to train deep bidirectional representations that consider both forward and backward contexts across all layers. In contrast, earlier language models were typically unidirectional. The pre-training phase of BERT in the context of Chinese humanities and social sciences consists of two main tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

The MLM task is a specialized auto-encoder training approach, where the BERT model randomly masks 15% of the input characters from Chinese humanities and social science texts and updates the parameters of the bidirectional transformer by predicting the masked portions. The NSP task requires the model to process sentence pairs and determine whether the two sentences belong to a coherent sequence, thereby learning inter-sentence relationships. By combining the loss functions from both tasks, BERT derives a federated learning loss function, enabling the simultaneous updating of the encoder's output layer and the connected classifier.

Pretraining data. The corpus consists of the abstracts in CSSCI (http://cssci.nju.edu.cn/login_u.html) and the full texts in China Social Science Excellence (http://ipub.exuezhe.com/index.html). We stored the crawled information in the MySQL database, from which we extracted all items whose values are not empty and obtained a total of 16,960,263 pieces of abstracts and paragraphs. Table 1 presents the descriptive statistical information of the corpus. The results indicate that the total number of data used in this experiment surpasses 4 billion Chinese characters.

In order to understand the distribution of disciplines in the Chinese humanity and social science Corpus, Tables 2 and 3 show the number distribution of different disciplines in abstracts and academic full-texts.

Experimental parameters for pre-training. Each line of the Chinese humanity and social science corpus is an abstract or paragraph of an article.Additionally, it's worth noting that the number of Chinese humanity and social science characters in the majority of lines is statistically less than 512. Therefore, for the pre-training experiment, the maximum sequence length is set to 512 to enhance the pre-training speed. The initial learning rate for the training of Chinese humanity and social science was set to 2e-5, and the warmup function provided by the transformers facilitated a rapid increase in the initial learning rate during the pre-training process for Chinese humanity and social science. Afterward, the learning rate gradually decreases to zero. This mode used in Chinese humanity and social science enabled the model to quickly stabilize during the initial training phase and subsequently converge rapidly in subsequent training sessions. Additionally, a decreasing learning rate struck a balance between the need for convergence speed and training accuracy. Considering the server configuration and model limitations, the training batch size values were individually set to 32 for each graphics card, and based on the usual experimental practice, three or five rounds of training were conducted to achieve better results. The specific parameter values are presented in Table 4.

Models performance evaluation

In this study, two approaches are used to evaluate the performance of pre-trained language models. The first method is simpler, assessing how well the model learns Chinese humanities and social science texts by calculating the model's perplexity on a test set. The second approach is more straightforward, applying the pretrained language model to a specific Chinese humanities and social science task. By observing the performance of the model in a Chinese humanities and social science task and comparing it with other models, its effectiveness can be assessed.

Evaluation of perplexity. The concept of perplexity introduces a novel solution. When dealing with substantial disparities in perplexity, a lower perplexity score indicates a stronger fit of the pre-trained model to real Chinese humanity and social science sentences. In other words, a lower perplexity signifies a superior model in Chinese humanity and social science. The computing formula for perplexity is in the following formula.

$$PP(W) = P\left(w_1 w_2 \ldots w_N\right)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P\left(w_1 w_2 \ldots w_N\right)}} \quad (1)$$

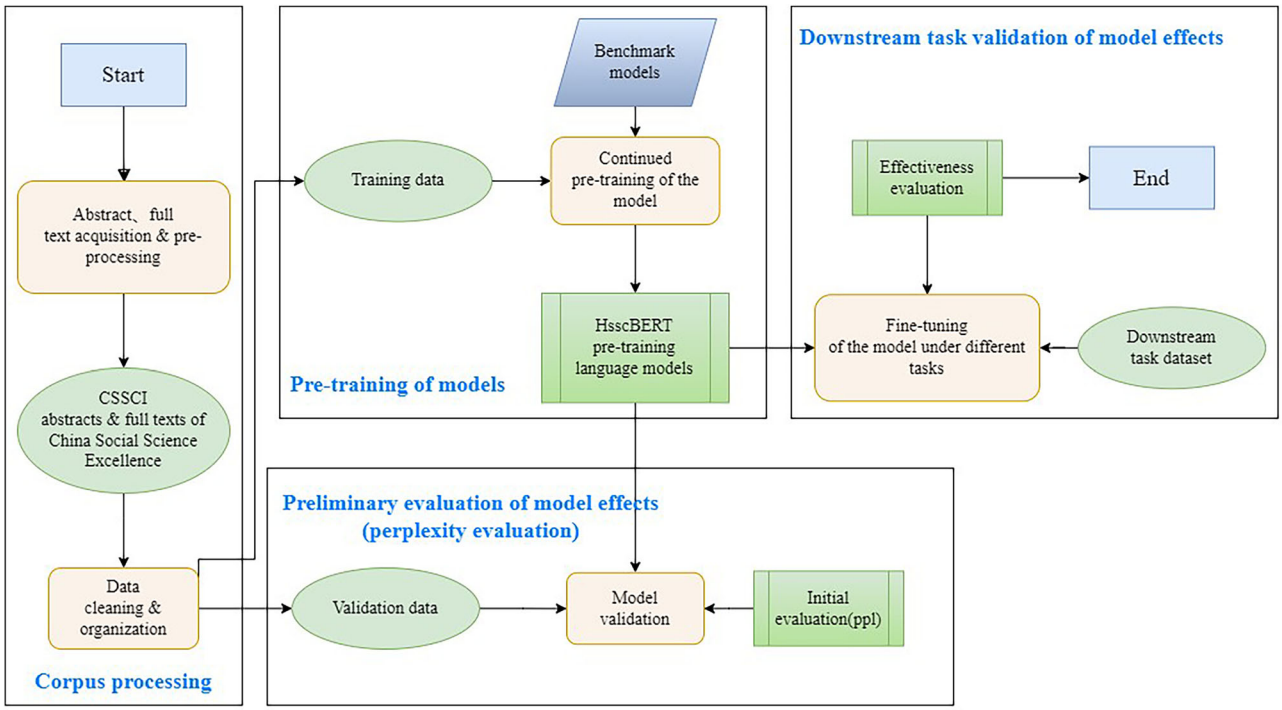


Fig. 1 | Framework of pre-training and validation experiments.

Table 1 | Chinese humanity and social science dataset

Basic statistical information	Values
Total file size	11.54 GB
Total number of abstracts and paragraphs	16,960,263
Total number of Chinese characters	4,285,801,244
No duplication of Chinese characters	19,856
Average number of Chinese characters each paragraph	252.7

During the experiment, several pre-trained experiments in Chinese humanity and social science were conducted, and the perplexities are presented in Table 5.

As can be observed in Table 5, two models pre-trained with corpus of Chinese humanity and social science have a low level of perplexity. Given that the test set for the experiments consists of abstracts or full academic text from actual publications, it can be assumed that the content of the test set is normal understandable sentences, and thus the language models with lower perplexity may perform better on the specific text mining task. Obviously, HsscBERT_e5 obtains a lower degree of perplexity than HsscBERT_e3, but it consumes much more time than HsscBERT_e3, and Hssc _ BERT _ e5 has better performance in specific natural language understanding tasks.

Evaluation of tasks in NLP. Perplexity, to some extent, reflects the convergence of the pre-trained model, and therefore, it can more effectively characterize the linguistic features of the pre-trained data. Indeed, while a model with lower perplexity indicates better understanding of the pre-trained data, it does not guarantee superior performance across various specific natural language understanding task. Thus, further experiments are needed to validate the performance of the model trained in Chinese humanity and social science. In this researches, a total of three natural language processing tasks were set for validation in the following three categories: (1) discipline classification experiments, namely, discipline classification tasks for titles and abstracts of CSSCI journal papers, (2) abstract structural function recognition experiments, including abstract sentence recognition tasks for articles published in *Data Analysis and Knowledge Discovery*, and (3) named entity recognition experiments, including Chinese literary datasets for entity recognition. The BERT-base-Chinese model and Chinese-RoBERTa-wwm-ext model were selected as the baseline models in this research for comparison with the model trained in Chinese humanity and social science. All of the above three types of tasks are important foundations for research in the field of academic textual studies in the humanities and social sciences, especially for in-depth analysis and research of literature. First, this study acquired and constructed a high-quality evaluation dataset, and divided the training set and test set according to 9:1. Second, the HsscBERT model and the selected BERT-like base model were fine-tuned and evaluated for their experimental results, respectively. Finally, the large language model is introduced to be tested on the test set.

Evaluation dataset. (1) Discipline classification dataset of CSSCI

The titles and abstracts of articles published in CSSCI-indexed journals from 1998 to the first half of 2021 were obtained and were categorized according to the CSSCI discipline classification standard. Approximately 2000 pieces of bibliographical information were randomly extracted from each subject, whereas all bibliographical information was extracted from subjects with fewer than 2000 papers. As such, bibliographical information was extracted from a total of 41,559 valid documents. The extracted bibliographical information was divided into a training and test set at a ratio of 9:1 in Chinese humanity and social science. The disciplines included in the dataset are shown in Table 6.

(2) Structural function recognition dataset

Data Analysis and Knowledge Discovery is one of the core CSSCI journals belonging to the discipline of library and information science. Relying on the fields of computer science, data science, and informetrics, among other fields, this journal extensively organizes research on the methods, theories, and technologies of data-driven knowledge discovery, intelligent management, and semantic computing. With the characteristics of both social and natural sciences, the journal is a product of big data management and application in the smart era. Because the abstracts of articles published in the journal have been strictly arranged in a structured

Table 2 | The distribution of CSSCI abstracts in accordance with discipline

Categories	Total with	%
Economics	197,188	13.43%
Management Science	152,346	10.38%
Education Science	121,309	8.26%
Political Science	108,418	7.38%
Science of Library, Information and Archival	104,068	7.09%
Science of Law	92,231	6.28%
Chinese Literature	85,817	5.84%
Philosophy	81,190	5.53%
History	80,560	5.49%
Art Theory	68,947	4.70%
Journalism and Communication	67,411	4.59%
Linguistics	50,828	3.46%
Sociology	46,311	3.15%
Science of Physical Culture and Sports	37,735	2.57%
Foreign Literature	24,052	1.64%
Psychology	22,603	1.54%
Marxism	19,728	1.34%
Environmental Science	19,315	1.32%
Cultural Science	18,796	1.28%
Ethnology	18,512	1.26%
Science of Religion	16,796	1.14%
Archaeology	14,806	1.01%
Human and economic geography	7498	0.51%
Statistics	5983	0.41%
Integrated Social Sciences	3309	0.23%
Military Science	2501	0.17%
Total	1,468,258	

Table 3 | The distribution of full texts from China Social Science Excellence in accordance with discipline

Categories	Total with	%
Economics and Economic Management	6,083,629	39.27%
Education	3,353,097	21.64%
Politics and Sociology	2,194,615	14.17%
Literature and Art	1,049,762	6.78%
Philosophy	794,990	5.13%
Dissemination of Cultural Information	656,319	4.24%
History	586,037	3.78%
Law	577,840	3.73%
Other	195,716	1.26%
Total	15,492,005	

form since 2014, the study of its abstract structural function recognition does not require a large number of manual tags, which avoids tag errors, and reduces the workload and time consumption.

In this dataset, abstract information of the literature published in the journal *Data Analysis and Knowledge Discovery* from 2014 to October 2021 (including articles first published online) was obtained. In addition, the abstracts were organized according to their labeled structured abstract subheadings ("Purpose," "Methods," "Results," "Limitations," and "Conclusions," among others) and classified into the corresponding categories and then normalized. Furthermore, the training and validation datasets of abstracts of *journal Data Analysis and Knowledge Discovery* were divided at a ratio of 9:1 with a single abstract as the unit of measurement. Table 7 presents the details of the structural function labels normalization.

(3) Chinese literary entity recognition dataset

To explore the effects of the pre-training of Hcss_BERT model, Chinese literary text data were selected for validation experiments in named entity recognition (NER)[37]. The data were obtained from a dataset for Chinese literary text entity recognition and relationship extraction, which was posted on GitHub (https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset/tree/master). The dataset is based on 726 articles related to Chinese literature filtered and extracted from the website. The dataset was constructed using an entity and relationship annotation, heuristic Tagging based on generic disambiguation rules, and CRF machine learning methods to assist in the tag and ensure the precision of the entities and relationships. The dataset was annotated using the "BIO" annotation method, which distinguishes entities from non-entities. There were seven entity categories in the dataset. In the entity categories, the dataset uses "Thing" to capture non-human objects, "Time" to capture story timelines, and "Metric" to capture the length and other information based on the characteristics of literary texts, allowing the label design to fit better with the text features (Table 8).

Table 4 | Pre-training parameters in Chinese humanity and social science

	HsscBERT_e3	HsscBERT_e5
max_seq_length	512	512
learning_rate	2e-05	2e-05
per_device_train_batch_size	32	32
per_device_eval_batch_size	16	16
num train epochs	3	5

Table 5 | Perplexity of the pre-trained models

Models	Benchmark model	Perplexity
BERT-base-Chinese	–	2.68
Chinese-RoBERTa-wwm-ext	–	2.81
HsscBERT_e3	BERT-base-Chinese	1.86
HsscBERT_e5	BERT-base-Chinese	1.83

Results

To verify the HcssBERT pre-trained model on various natural language processing tasks, we conducted comparative experiments between the proposed and following baseline models. Based on the models used during pre-training described in the previous section, the baseline models used in experiments are as follows: (1) BERT-base-Chinese, the basic Chinese pre-trained model officially provided by Google, was selected as the first validation baseline model. (2) Chinese-RoBERTa-wwm-ext, the Chinese RoBERTa model adopting. (3) ChatGPT, one of the top-performing generative multilingual pre-trained models at the current stage, we utilized GPT-3.5-turbo and GPT-4.0-turbo from the GPT series to provide response services. (4) LLAMA3.1-8B, one the best-performing open-source large language models, for LLAMA3.1-8B model, this study firstly performs Lora fine-tuning, and then uses the fine-tuned model for downstream task validation. In light of the common performance metrics system for NLP, in this pre-trained model of Chinese humanity and social science validation task, we used the Accuracy, Precision (P), Recall (R), F1-score, Macro-avg, and

Table 6 | Discipline classification labels and data distribution

Academic discipline	Number of studies	Effective documentation	Training set	Test set	Academic discipline	Number of studies	Effective documentation	Training set	Test set
Science of law	10,284	7192	1796	204	Library, Information Science and Documentation	6594	5274	1822	178
Management science	13,513	8558	1814	186	Foreign literature	1749	1214	1103	111
Education science	12,758	9199	1810	190	Psychology	2052	1832	1642	190
Economics (as a field of study)	33,030	17,404	1803	197	Journalism and Communication	3800	1794	1624	170
Archaeology	421	110	100	10	Art Theory	3869	1638	1471	167
Rare studies	490	411	366	45	Linguistics	2523	1802	1621	181
History	9311	5959	1795	205	Philosophy	7994	4239	1777	223
Marxist theory	11,423	7038	1782	218	Political Science	13,901	9354	1799	201
Ethnography and cultural studies	1632	1234	1091	143	Chinese literature	6068	2032	1805	195
Human and economic geography	3074	2472	1782	218	Natural Resources and Environmental Science	950	714	645	69
Sociology	3353	2013	1807	193	Science of Religion	560	409	377	32
Science of physical culture and sports	2102	1779	1608	171	Comprehensive journal sponsored by universities	2790	2021	1795	205
Statistics	1038	622	559	63	Integrated social sciences	37,241	24,811	1809	191

Table 7 | Uniform specifications of structural function labels

Normalized structural function labels	Structural function labels in the original article
Purpose	Purpose, objectives, context of application
Methodology	Methodology, scope of literature
Findings	Findings
Limitations	Limitations
Conclusion	Conclusion

Weighted-avg to evaluate the experimental performance of the four pre-trained models. Each metric is calculated as follows (Table 9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}Score = \frac{2precision * recall}{precision + recall} \tag{5}$$

The macro-average in Chinese humanity and social science represents the arithmetic mean of the statistical indicators for all categories, including macro-accuracy, macro-recall, and macro-F1-score.

$$macro\text{-}precision = \frac{1}{n}\sum_{i=1}^{n} precision_i \tag{6}$$

$$macro\text{-}recall = \frac{1}{n}\sum_{i=1}^{n} recall_i \tag{7}$$

$$macro\text{-}F1\text{-}score = \frac{2precison_{macro} * recall_{macro}}{precison_{macro} + recall_{macro}} \tag{8}$$

The weighted-average in Chinese humanity and social science involves selecting the ratio of the number of samples in each category to the total number of samples in all categories as weights, and then computing the average. The specific indicators to be calculated using the weighted-average in Chinese humanity and social science are the weighted precision, weighted recall, and weighted F1 values.

$$weighted\text{-}precision = \sum_{i=1}^{n} precision_i * f_i \tag{9}$$

$$weighted\text{-}recall = \sum_{i=1}^{n} recall_i * f_i \tag{10}$$

$$weighted\text{-}F1\text{-}score = \frac{2precison_{weighted} * recall_{weighted}}{precison_{weighted} + recall_{weighted}} \tag{11}$$

Prompts for LLM

Table 10 presents the prompts used for three different downstream tasks. In the case of text classification tasks like structural function recognition and discipline classification, no additional samples were added to the prompts which relied solely on the provided descriptions. However, for the named entity recognition (NER) task, more complex prompts were constructed,

including a specific example. This is because NER tasks have a higher degree of output discreteness compared to classification tasks, making it challenging to extract computational elements from generated text without structured control over the output. Nonetheless, the model still has the possibility of generating text outside the label set. To address this issue, during the final calculation process, any output generated by the model that does not belong to any label category is regarded as an erroneous label, ensuring the rigor of the computed results.

Experiment results

Discipline classification experiments. In this study, the baseline models and the two proposed pre-trained models for Chinese humanity and social science domains were used to conduct automatic classification experiments based on the title and abstract text of CSSCI journal articles belonging to the disciplines, respectively. The classification of the models was measured using the accuracy, macro-avg, and weighted-avg values. Table 10 presents the results of the classification experiments. The data indicates that HsscBERT_e3 proposed in this study achieved the optimal results on the discipline classification experiments conducted on the title, and the best accuracy and weighted avg value in the combined data of abstract. HsscBERT_e5 works best for the macro average of the abstract. GPT performs poorly in two scenarios, with a low macro-average. This is due to the unpredictability of the generated content by the generative model, resulting in the generation of data beyond the label set. Compared with the two models using WWM technology, HsscBERT_e3 and HsscBERT_e5 outperformed the baseline model, indicating that the multiple models proposed in this study can provide better support for the intelligent processing of humanity and social science texts. It is worth noting that HsscBERT_e3 is slightly better than HsscBERT_e5 in the title classification task, which should be related to the characteristics of the training data, and the full-text data accounted for a relatively large proportion of the data used in this research, so HsscBERT_e3 achieved better results in the shorter text task. The GPT3.5-turbo model performs weakly, especially at the macro-average. The reason for this is that the GPT3.5-turbo model often outputs redundant content and content that is not in the given category set, which seriously affects the performance of the model. In contrast, the performance of the GPT4-turbo model has greatly improved. Although there is still a certain gap compared to the fine-tuned model, it has achieved an effect close to the fine-tuned BERT model in the 0-shot situation. Meanwhile, it is evident that the performance improvement of the GPT4-turbo model is very significant when the input content is richer. In addition, the LLAMA3.1-8B model after lora fine-tuning outperforms GPT4.0-turbo on the classification task, but also suffers from irregularities in the output format, where the output does not belong to the given category.

Structural function recognition experiments. To verify the performance of the pre-trained model in Chinese humanity and social science, abstract sentence structural function recognition experiments were conducted using abstracts obtained from publications in *Data Analysis and Knowledge Discovery*. The abstracts of the journal are standardized and can be directly used as standard answers for structural function recognition after a simple normalization. The experiments in in Chinese humanity and social science used the precision, recall, and the F1-score under macro-avg and weighted-avg as the basic metrics to evaluate the models performance. The results were compared with BERT-base-Chinese, Chinese-Roberta-wwm-ext and GPT. The model validation results are shown in Table 11. The experiments shows that HsscBERT_e5 has achieved the optimal value in every index. In the meantime, the performance of HsscBERT_e3 model is slightly lower than that of HsscBERT_e5. In contrast, the Chinese-Roberta-wwm-ext model performs the poorer in this task, with all indexes not reaching 80%, while the performance of BERT-base-Chinese, one of the baseline models, is fair medium performance. The ability of GPT3.5-turbo to recognize abstract sentence structural function is poorest, but compared to subject classification tasks, GPT3.5-turbo achieves higher macro-average scores in sentence segmentation tasks. It generates very little text outside the generated label set, indicating that large language models tend to produce more structured text when performing tasks with a smaller number of

Table 8 | Entity labels of dataset for Chinese literary entity recognition

Label	Description	Proportion
Thing	Thing	35.63
Person	Person	32.71
Location	Location, country, or city name	17.18
Time	Time related words	7.36
Metric	Measurement related words	3.64
Organization	Organization name	2.03
Abstract	Abstract	1.43

Table 9 | Confusion matrix of models in Chinese humanity and social science

Predicted class		
Class	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

Table 10 | Input prompt examples for LLM

Task name	Prompt
Discipline classification	Do not output anything else, just give the tag name, please choose the most appropriate one from the given tag set ['history', 'sports science', 'economics', 'management', 'psychology', 'arts', 'archaeology', 'law', 'natural resources and environmental science', 'ethnology and cultural studies', 'sociology', 'statistics', 'Marxist theory', 'humanistic economic geography', 'obscure extinct discipline', 'library, information and literature science', 'education', 'Chinese literature', 'foreign literature', 'comprehensive social science', 'linguistics', 'philosophy', 'journalism and communication', 'political science', 'religious studies', 'comprehensive university journal'] to classify the following scientific text, only one tag from the given tag set can be assigned to the given text, the given text is: {}
Structural function recognition	Do not output anything else, just give the tag name, please choose the most appropriate tag from the given tag set ['conclusion','purpose','results','method','limitations'] to classify the following text, the given text is: {}
Named entity recognition	Just output the annotated result, the given tag set is ['Person','Location','Time','Organization','Metric','Thing','Abstract'], where Person represents names, pronouns, or titles, Location represents place names, Organization represents organization names, Time represents temporal nouns, Metric represents quantifiers other than temporal nouns, Thing represents ordinary objects, and Abstract represents abstracts. Please use these tags to annotate the given text. The output should retain the original text and annotate entities with "<tag>entity</tag>", for example, if the given text is: "In 1935, a never-before-seen flood of evil water flattened all the buildings on the land of Dongziping to the ground and formed pits, devouring 176 lives." then you should annotate it as: "<Time>1935</Time>, a never-before-seen flood of <Thing>evil water</Thing> flattened all the buildings on the land of <Thing>Dongziping</Thing> to the ground and formed <Thing>pits</Thing>". The text you need to process is: {}

Table 11 | Experiment results of classification models

	Title			Abstract		
	Accuracy (%)	Macro Avg (%)	Weighted Avg (%)	Accuracy (%)	Macro Avg (%)	Weighted Avg (%)
BERT-base-Chinese	65.28	61.78	63.42	70.57	67.65	69.11
Chinese-Roberta-wwm-ext	58.66	52.89	56.12	64.82	58.98	62.37
HsscBERT_e3	**65.74**	**61.95**	**63.75**	**71.63**	68.69	**70.09**
HsscBERT_e5	65.47	61.83	63.52	71.51	**68.94**	69.98
LLAMA3.1-8B	61.42	56.82	61.22	64.01	60.90	62.39
GPT3.5-turbo	42.56	4.08	42.14	44.90	8.30	43.93
GPT4-turbo	62.58	41.36	60.87	63.54	43.27	63.13

Table 12 | Experiment results of structural function identification of abstracts (%)

	Accuracy (%)	Macro Avg (%)	Weighted Avg (%)
BERT-base-Chinese	88.00	87.71	88.00
Chinese Roberta-wwm-ext	79.83	79.24	79.82
HsscBERT_e3	88.35	88.09	88.34
HsscBERT_e5	**88.52**	**88.32**	**88.54**
LLAMA3.1-8B	85.86	86.23	85.39
GPT3.5-turbo	54.26	44.00	54.12
GPT4-turbo	75.64	70.23	72.83

Table 13 | Experimental results of Chinese literary entity identification (%)

	Precision (%)	Recall (%)	F1 (%)
BERT-base-Chinese	69.51	72.53	70.99
Chinese-RoBERTa-wwm-ext	48.37	52.95	50.56
HsscBERT_e3	**71.31**	73.45	72.36
HsscBERT_e5	71.17	**73.64**	**72.38**
LLAMA3.1-8B	67.32	64.15	65.70
GPT3.5-turbo	43.82	15.43	18.20
GPT4-turbo	60.46	32.87	42.59

labels. The GPT4 turbo model has made significant improvements at the macro coverage level, partly due to the superior performance of the GPT4 turbo model itself. On the other hand, the categories involved in abstract step recognition tasks are far less than those in text subject classification, which also enables the model to better predict categories in 0-shot situations. Compared to the subject classification task, the LLAMA3.1-8B model performs even better on the speech step recognition task, with much better performance than the GPT family of models, and there is no case of outputting a non-given set of categories due to fewer data categories.

Chinese literary entity recognition. As one of the tasks of NLP, named entity recognition is the basis for subsequent applications such as question answering systems, machine translation, and information retrieval. To verify the recognition effect of the HsscBERT pre-trained models, the Chinese literature dataset was used for entity recognition validation experiments, which contained seven types of entities including 'characters' and 'places'. The precision, recall, and F1-score were used as the evaluation indexes of the model as is typical with named entity recognition. The experimental results are presented in Table 12. It can be seen that HsscBERT_e5, obtained the optimal performance in both Recall and F1, and the F1-score is 73.38% for all types of entities, which is 1.39% higher than the benchmark model BERT-base-Chinese, and poor performance Chinese_RoBERTa_wwm_ext, only reaching 50.56%. The performance of GPT3.5-turbo on this task is also lower than that of the fine-tuned model, and the recall is significantly lower than the precision. The F1 score only reaches 18.20%. This indicates that even the currently most advanced model is still difficult to apply to vertical domain tasks. Similar to the previous two experiments, GPT4-turbo still exhibits much better performance than the GPT3.5 turbo model, but the macro coverage index is still low. Obviously, in the 0-shot scenario, it is difficult to require generative models such as GPT to complete domain tasks while ensuring the standardization of the output. The LLAMA3.1-8B model outperforms the GPT model on the named entity recognition task, but there is still a gap compared to the BERT family of models, and in combination with the other two validation results, the generative model still needs to be explored more on the text comprehension-type task (Table 13).

Discussion

The present study focuses on the development of pre-trained language models specifically for Chinese humanities and social science texts. A series of models was proposed, including HsscBERT_e3 and HsscBERT_e5, and their performance was evaluated across a range of tasks, including discipline classification, abstract structural function recognition, and knowledge entity recognition. The models demonstrated superior performance in domain-specific tasks in comparison to baseline models, particularly with regard to the recognition of domain-related linguistic features and the improvement of semantic understanding. Furthermore, experimental findings indicated that five rounds of pre-training yielded superior model performance in comparison to three rounds, thereby underscoring the significance of continuous domain-specific training. Additionally, the research highlighted that small and fine-tuned models outperform large-scale models like GPT in certain specialised tasks, affirming the value of targeted model training over broad, generic pre-training. The process described in this study can be further generalized to other domains, such as literature, history and philosophy. Training on large-scale domain texts enables the model to demonstrate superior performance on domain-specific tasks.

Notwithstanding the favourable outcomes, this study also identified certain limitations. From a cross-disciplinary and cross-language perspective, the performance of the model varies across disciplines due to differences in publication volume and language style. The more esoteric language of disciplines such as philosophy poses a significant challenge for training, leading to lower recognition accuracy in specific tasks. The increasing overlap of disciplinary content in modern research also complicates text categorisation, as it makes it more difficult for models to achieve high classification accuracies. While the model has demonstrated proficiency in Chinese humanities and social sciences, challenges persist in multilingual scenarios due to constraints in model parameters and training data. From the perspective of NLP tasks, the encoder-only architecture model represented by BERT is more suitable for text comprehension tasks, and is difficult to perform generative tasks (e.g., translation, summarization).

To date, NLP has experienced four research paradigms: Fully Supervised Learning (Non-Neural Network), Fully Supervised Learning (Neural Network), Pre-train, Fine-tune and Pre-train, Prompt, Predict[38]. This study represents an attempt and exploration of the third paradigm in the context of Chinese humanities and social sciences. The experimental results demonstrate that, despite the large language model's demonstrated proficiency in text generation within the general domain, the Pre-train, Fine-tune research paradigm based on small models maintains relevance in the vertical domain. Future research should address the limitations of perplexity as a metric for model evaluation, exploring alternative metrics that could better capture nuanced performance differences, especially in low perplexity scenarios. In addition, further exploration is required to enhance the performance of models across diverse disciplinary domains characterised by varied linguistic characteristics. The incorporation of a more diverse array of data sources, along with the identification and resolution of challenges pertaining to cross-disciplinary classification, will serve to enhance the model's versatility and generalizability. Additionally, refining the pre-training process to accommodate the specific needs of different linguistic styles, such as those found in Chinese humanities and social sciences, could further optimize model performance. The exploration of advanced techniques, such as the incorporation of external knowledge bases, the integration of more domain-specific features, and the application of hybrid models combining supervised and unsupervised learning approaches, holds promise for the enhancement of future model development. As the field of natural language processing (NLP) continues to evolve, it is essential to examine how new paradigms, such as the fourth paradigm of NLP, can be integrated into the pre-training process to improve performance on downstream tasks.

Data availability

Access to the data is available at https://github.com/S-T-Full-Text-Knowledge-Mining/HsscBERT.

Received: 22 September 2024; Accepted: 28 April 2025;
Published online: 03 June 2025

References

1. Otter, D. W., Medina, J. R. & Kalita, J. K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 604–624 (2020).
2. Lauriola, I., Lavelli, A. & Aiolli, F. An introduction to deep learning in natural language processing: models, techniques, and tools. *Neurocomputing* **470**, 443–456 (2022).
3. Goldberg, Y. *Neural Network Methods for Natural Language Processing* (Springer Nature, 2022).
4. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
5. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013).
6. Pennington, J., Socher, R. & Manning, C. D. Glove: global vectors for word representation. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014).
7. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).
8. Devlin, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics 4171–4186 (NAACL, 2019).
9. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30* (NIPS, 2017).
10. Yu, T. R. et al. Review of research on pre-training models for natural language processing. *Comput. Eng. Appl.* **56**, 12–22 (2020).
11. Qiu X. et al. Pre-trained models for natural language processing: a survey. *Sci. China Technol. Sci.* **63**, 1872–1897 (2020).
12. Hu, D. et al. A multilingual text sentiment analysis method incorporating attention mechanism under pre-training model. *J. Chin. Comput. Syst.* **02**, 278–284 (2020).
13. Chen, S., Pei, Y., Ke, Z. & Silamu, W. Low-resource named entity recognition via the pre-training model. *Symmetry* **13**, 786 (2021).
14. Caciularu, A. et al. CDLM: Cross-Document Language Modeling. In Findings of the Association for Computational Linguistics 2648-2662 (EMNLP, 2021).
15. Xu, Y. et al. Layoutlm: pre-training of text and layout for document image understanding. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 1192–1200 (ACM, 2020).
16. Luo, Z., Xi, Y., Zhang, R. & Ma, J. VC-GPT: visual conditioned GPT for end-to-end generative vision-and-language pre-training. Preprint at https://arxiv.org/abs/2201.12723 (2022).
17. Li, X., Han, D. & Chang, C. C. Pre-training model based on parallel cross-modality fusion layer. *PLoS ONE* **17**, e0260784 (2022).
18. Dai, Z. et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 2978–2988 (ACL, 2019).
19. Yang, Z. et al. Xlnet: generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32* (NIPS, 2019).
20. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T. Y. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems* 16857–16867 (NIPS, 2022).
21. Ding, M., Zhou, C., Yang, H. & Tang, J. Cogltx: applying BERT to long texts. In *Advances in Neural Information Processing Systems* 12792–12804 (NIPS, 2020).
22. Jiao, X. et al. TinyBERT: Distilling BERT for Natural Language Understanding. Findings of the Association for Computational Linguistics 4163–4174 (EMNLP, 2020).
23. Xu, C. et al. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing 7859–7869 (EMNLP, 2020).
24. He, R., Sun, S., Yang, J., Bai, S. & Qi, X. Knowledge distillation as efficient pre-training: faster convergence, higher data-efficiency, and better transferability. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9161–9171 (IEEE, 2022).
25. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
26. Jeong, M. et al. Transferability of Natural Language Inference to Biomedical Question Answering. CEUR Workshop Proceedings, 2696 (2020).
27. Xiao, Q., Zhou, X., Xiao, Y. & Zhao, K. Yunnan university at vqa-med 2021: pretrained bioBERT for medical domain visual question answering. In *Working Notes of CLEF* 201 (2021).
28. Liu, F. et al. Self-Alignment Pretraining for Biomedical Entity Representations. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics 4228–4238 (NAACL, 2021).
29. Hebbar, S. & Xie, Y. CovidBERT-biomedical relation extraction for Covid-19. In *The International FLAIRS Conference Proceedings* 34 (FLAIRS, 2021).
30. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, 86 (2021).
31. Beltagy, I. et al. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing 3615–3620 (EMNLP-IJCNLP, 2019).
32. Park, S. & Caragea, C. Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In *Proc. 28th International Conference on Computational Linguistics* 5409–5419 (2020).
33. van Dongen, T., Wenniger, G. M. D. B. & Schomaker, L. SChuBERT: scholarly document chunks with BERT-encoding boost citation count prediction. Preprint at https://arxiv.org/abs/2012.11740 (2020).
34. LLiu, X. et al. OAG-BERT: Towards a Unified Backbone Language Model for Academic Knowledge Services. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 3418–3428 (KDD, 2022).
35. Shen, S. et al. SsciBERT: a pre-trained language model for social science texts. *Scientometrics* **128**, 1241–1263 (2023).
36. Li, Y. et al. CSL: a large-scale Chinese scientific literature dataset. In *Proceedings of the 29th International Conference on Computational Linguistics* 3917–3923 (ACL, 2022).
37. Xu, J., Wen, J., Sun, X. & Su, Q. A discourse-level named entity recognition and relation extraction dataset for Chinese literature text. Preprint at https://arxiv.org/abs/1711.07010 (2022).
38. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).

Acknowledgements

We would like to express my sincere thanks to the data annotation team of Nanjing Agricultural University. This work was supported the National Natural Science Foundation of China (Grant Number: 71974094).

Author contributions

Si Shen designed the research and was responsible for writing the paper. Zhixiao Zhao completed the model parameter tuning. Dayu Yan cleaned and organized the data. Chuan Jiang conducted experimental validation with large language models. Dongbo Wang designed the paper framework.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Si Shen.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025