Surveying the Dead Minds: Historical-Psychological Text Analysis with Contextualized Construct Representation (CCR) for Classical Chinese

Anonymous ACL submission

Abstract

In this work, we develop a pipeline for historical-psychological text analysis in classical Chinese. Humans have produced texts in various languages for thousands of years; however, most of the computational literature is focused on contemporary languages and corpora. The emerging field of historical psychology relies on computational techniques to extract aspects of psychology from historical corpora using new methods developed in natural language processing (NLP). The present pipeline, called Contextualized Construct Representations (CCR), combines expert knowledge in psychometrics (i.e., psychological surveys) with text representations generated via transformer-based language models to measure psychological constructs such as traditionalism, norm strength, and collectivism in classical Chinese corpora. Considering the scarcity of available data, we propose an indirect supervised contrastive learning approach and build the first Chinese historical psychological corpus (C-HIS-PSY) to fine-tune pre-trained models. We evaluate the pipeline and benchmark it against objective external data to test its validity. We also release our dataset and code for reproducibility at https://anonymous.4open.science/r/His-Psy/.

1 Introduction

Humans have been producing written language for thousands of years. Historical populations have expressed their norms, values, stories, songs, and more in these texts. Such historical corpora represent a rich yet underexplored source of psychological data that contains the thoughts, feelings, and actions of people who lived in the past (Jackson et al., 2021). The emerging field of "historical psychology" has been developed to understand how different aspects of psychology vary over historical time and how the origins of our contemporary psychology are rooted in historical processes (Atari and Henrich, 2023; Muthukrishna et al., 2021; Baumard et al., 2024). Since we cannot access "dead minds" directly but can access their textual remains, natural language processing (NLP) is the primary method to extract aspects of psychology from historical corpora. Previous works are often monolingual and in English (Blasi et al., 2022). In addition, much of the literature at the intersection of psychology and NLP has relied on bag-of-words or word embedding models, focusing on non-contextual word meanings rather than a holistic approach to language modeling.

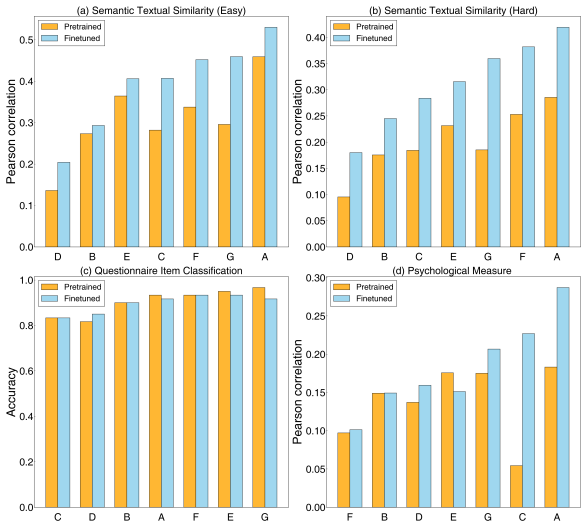

Figure 1: Comparison of model performance on the three tasks in the test set before and after fine-tuning. (Model A: bert-ancient-chinese, B: guwenbert-base, C: guwenbert-large, D: paraphrase-multilingual-MiniLM-L12-v2, E: text2vec-base-chinese, F: text2vec-base-chinese-paraphrase, G: text2vec-large-chinese)

Recently, more research attention in the NLP community has been directed to historical and ancient languages (Johnson et al., 2021), including but not limited to English (Manjavacas Arevalo and Fonteyn, 2021), Latin (Bamman and Burns, 2020), ancient Greek (Yousef et al., 2022), and ancient Hebrew (Swanson and Tyers, 2022). While all

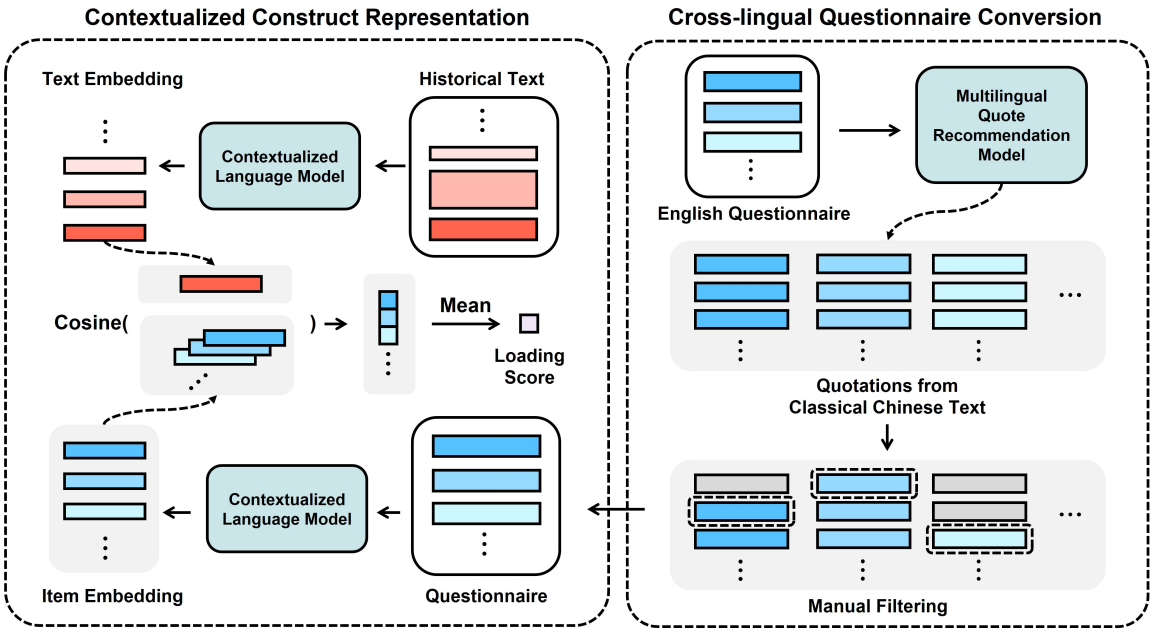

Figure 2: Pipeline of cross-lingual questionnaire conversion and contextualized construct representation for classical Chinese.

these languages have historical significance, classical Chinese is particularly important in the quantitative study of history. China has a long history spanning thousands of years, largely recorded in classical Chinese. The language served as a medium for expressing and disseminating influential philosophical and religious ideas. Confucianism, Daoism, and later Buddhism (through translations from Sanskrit) all found expression in classical Chinese, profoundly shaping Chinese thought, ethics, governance, and norms. As more resources become readily available for classical Chinese, scholars of ancient China can test more specific hypotheses using computational methods (Liu et al., 2023; Slingerland, 2013; Slingerland et al., 2017).

Due to its historical significance and geographical coverage, classical Chinese represents one of the most important languages in the study of historical psychology (Atari and Henrich, 2023). Prior work in social science has often relied on bag-of-words approaches (Zhong et al., 2023) or bottom-up techniques such as topic modeling (Slingerland et al., 2017). In the NLP community, while different Transformer-based models of classical Chinese have been developed, they have not been applied to theory-driven psychological text analysis. For example, AnchiBERT is a specialized pre-training model tailored to analyze classical Chinese literature (Tian et al., 2021). Its pre-training data consists of 39.5 million tokens from ancient Chinese, covering many texts, such as historical documents, essays, classical poetry, and verses, over millennia. Employing AnchiBERT for generating Ancient Chinese content involves leveraging a Transformer-based architecture (Vaswani et al., 2017). This model has been used in NLP tasks such as translation (Wang et al., 2023). Still, they have not been used to extract psychological constructs (e.g., moral values, norms, cultural orientation, mental health, religiosity, emotions, and thinking styles) from historical corpora. Transformer-based language models are crucial for psychological text analysis because psychological constructs are often complex, and sentence-level semantics (and above) will more effectively capture psychological meanings than isolated words (Demszky et al., 2023) or non-contextual word embedding models.

Here, we create a pipeline called Contextualized Construct Representation (CCR) for historical-psychological text analysis in classical Chinese. Although CCR has recently been developed for contemporary psychological text analysis (Atari et al., 2023b), it can be adapted for historical NLP because it relies on Transformer-based models. As a tool for psychological text analysis, CCR takes advantage of contextual language models in NLP, does not require selecting a priori lists of words to represent a psychological construct (e.g., the popular Linguistic Inquiry and Word Count program, Boyd et al., 2022), and takes advantage of

psychometrically validated questionnaires in psychology. CCR proceeds in four steps: (1) selecting a questionnaire for the psychological construct of interest; (2) representing questionnaire items as embeddings using a contextual language model; (3) generating the embedding of the target text using a contextual language model; (4) computing the cosine similarity between the item and text embeddings. This straightforward pipeline is particularly useful for social science, wherein researchers are interested in interpretability and hypothesis testing. Previous work has shown that CCR outperforms other top-down methods such as dictionaries (Atari et al., 2023b), can replicate prior findings and similar methods (Simchon et al., 2023), and performs similarly to Large Language Models (LLMs) such as ChatGPT for psychological text annotation (Abdurahman et al., 2023).

2 Related Work

Psychological Text Analysis Given the increasing amount of online textual data, many social scientists are turning to NLP to test their theories. Unlike in some computational fields, social scientists traditionally give primacy to "theory" rather than prediction (Yarkoni and Westfall, 2017). Hence, theory-driven text analysis is the first methodological choice in social sciences, including psychology (Jackson et al., 2021; Wilkerson and Casas, 2017; Boyd and Schwartz, 2021). Given the importance of theory development and hypothesis testing, many social scientists have developed dictionaries to assess psychological constructs as diverse as moral values (Graham et al., 2009), stereotypes (Nicolas et al., 2021), polarization (Simchon et al., 2022), and threat (Choi et al., 2022).

Distributed Dictionary Representation (DDR) Aiming to integrate psychological theories with the capabilities of word embeddings, Garten et al. (2018) proposed the Distributed Dictionary Representation (DDR) as a top-down psychological text-analytic method. This method involves (a) defining a concise list of words by social scientists to capture a specific concept, (b) using a word-embedding model to represent these individual words, (c) computing the centroid of these word representations to define the dictionary's representation, (d) determining the centroid of the word embeddings within a given document, and (e) assessing the cosine similarity between the dictionary's representation and that of the document. DDR has been a useful approach in measuring moral and political rhetoric (Wang and Inbar, 2021), temporal trends in politics (Xu et al., 2023), and situational empathy (Zhou et al., 2021).

Contextualized Construct Representation (CCR) The Contextualized Construct Representation (CCR) (Atari et al., 2023b) pipeline is built upon SBERT (Reimers and Gurevych, 2019). This theory-driven and flexible approach has been shown to outperform dictionary-based methods and DDR for various psychological constructs such as religiosity, moral values, individualism, collectivism, and need for cognition. Furthermore, recent work suggests that CCR performs on par with LLMs such as GPT4 in measuring psychological constructs (Abdurahman et al., 2023). Although CCR has not been developed specifically for historical psychology, its flexible pipeline and easy-to-implement steps offer a unique opportunity to extract psychological constructs from historical corpora. In a way, CCR is similar to DDR, but instead of relying on non-contextual word embeddings, it makes use of the power of contextual language models to represent whole sentences (or larger texts). In addition, it obviates the development of word lists; instead, making use a thousands of existing questionnaires that have been validated in psychology over the last century.

Semantic Textual Similarity While BERT (Devlin et al., 2018) can identify sentences with similar semantic meanings, this process can be resource-intensive. To enhance the performance of BERT for tasks like semantic similarity assessments, clustering, and semantic-based information retrieval, Reimers and Gurevych (2019) developed Sentence-BERT (or S-BERT). This model employs a Siamese network structure specifically designed to create embeddings at the sentence level. S-BERT outperforms conventional transformer-based models in tasks related to sentences and significantly reduces the time needed for computations. It is engineered to generate sentence embeddings that capture the core semantic content, ensuring that sentences with comparable meanings are represented by closely positioned embeddings in the vector space. Therefore, S-BERT provides an efficient and less computationally demanding method for evaluating semantic similarities between sentences, making it particularly useful in fields such as psychology (Juhng et al., 2023; Sen et al., 2022).

3 Methodology

3.1 Cross-lingual Questionnaire Conversion

The process of converting a contemporary English questionnaire $\mathcal{Q}$ into a classical Chinese questionnaire $\tilde{\mathcal{Q}}$ is illustrated in the right panel of Figure 2. For each questionnaire item ($q_i \in \mathcal{Q}$), the multilingual quote recommendation model, "QuoteR" (Qi et al., 2022), which is trained on a dataset that includes English, modern Standard Chinese, and classical Chinese, can identify a set of quotations $\{\tilde{q}\}_i$ in classical Chinese that are semantically similar to the English sentence q_i.

For each questionnaire, all the items were entered into the model, resulting in a pool of corresponding quotations. A manual filtering process followed this to eliminate quotations of low quality, which can be either inappropriate or irrelevant to the psychological construct. Ultimately, the most similar quotations $\tilde{q}_i$ were selected, substituting for every English q_i to construct $\tilde{\mathcal{Q}}$ in classical Chinese.

3.2 Indirect Supervised Contrastive Learning

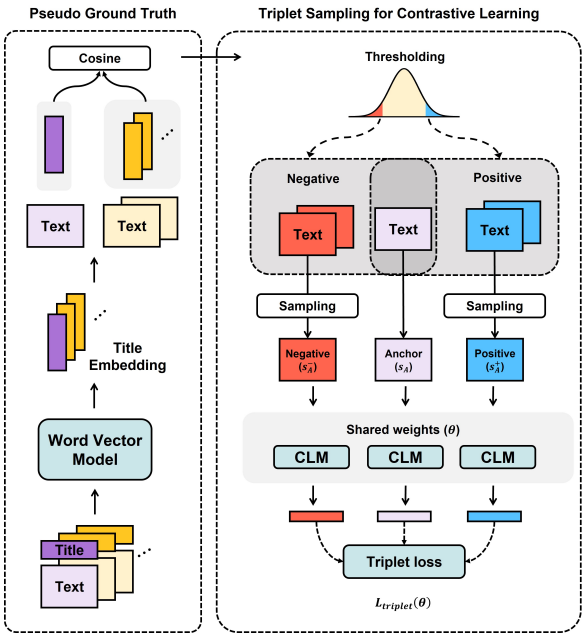

Figure 3: Pipeline of triplet sampling and contrastive learning. CLM stands for contextualized language model.

To obtain better psychology-specific CCR for Chinese historical texts, we introduce an indirect supervised contrastive learning approach to fine-tune pre-trained sentence embedding models, as shown in Figure 3.

Historical Psychology Corpus We assemble a refined corpus named Chinese historical psychology corpus (C-HIS-PSY) (https://anonymous.4open.science/r/His-Psy/dataset/), which is comprised of 21,539 paragraphs ($\mathcal{S}$) extracted from 667 distinct historical articles and book chapters in classical Chinese. The titles of these works ($\mathcal{T}$, $|\mathcal{T}| \ll |\mathcal{S}|$), each carefully selected for their relevance to moral values, serve as labels for their topics, including "節義" (moral integrity), "孝弟" (filial piety and fraternal duty), "盡忠" (utmost loyalty), "廉恥" (sense of shame), "清介" (pure and incorruptible), and "愛己" (love oneself).

We divide our data into training, validation, and testing sets, allocating 60%, 20%, and 20% of the data to each set, respectively. The distribution of paragraph lengths across different sets is consistent, as shown in Figure 6.

Pseudo Ground Truth from Titles Since the title ($t_i \in \mathcal{T}$) of a paragraph ($s_i \in \mathcal{S}$) is a concise summary of the moral values reflected in the paragraph, the semantic similarity between titles, $\textbf{sim}(t_i, t_j)$, can be considered as the pseudo ground truth for the semantic similarity between corresponding paragraphs, $\textbf{sim}(s_i, s_j)$. The semantic similarity between titles can be obtained by embedding the titles via $E_T(\cdot)$ and calculating their cosine similarity $\textbf{cos}(E_T(t_i), E_T(t_j))$. To perform word embedding on the titles, We trained five word vector models on a large classical Chinese corpus containing over a billion word tokens using different frameworks and architectures, and picked the best-performing one (see Appendix B for word vector model details).

Positive and Negative Sampling We calculate the cosine similarities between the title embeddings $\textbf{cos}(E_T(t_i), E_T(t_j))$, obtained through the word vector model, of all title pairs (the Cartesian product $\mathcal{T} \times \mathcal{T}$) in the corpus. The distribution of title similarities is illustrated in Figure 7. We obtain positive and negative paragraph pairs by thresholding the similarities of title pairs. Paragraphs whose titles have similarities exceeding the upper threshold δ^+, as well as those with identical titles, were identified as positive pairs $(\mathcal{S} \times \mathcal{S})^+$, that is,

$$\{(s_i, s_j)^+ \mid \textbf{sim}(E_T(t_i), E_T(t_j)) > \delta^+\}$$

Conversely, those with titles having similarities below the lower threshold δ^- were designated as

negative pairs $(\mathcal{S} \times \mathcal{S})^-$, that is,

$$\{(s_i, s_j)^- \mid \textbf{sim}(E_T(t_i), E_T(t_j)) < \delta^-\}$$

We experiment with several threshold settings, including 0.5th/99.5th, 1st/99th, 10th/90th, and 25th/75th percentiles. Our findings demonstrate that the 10th/90th percentile threshold yields the best performance, see Figure 4. Hence, for the following experiments, if not specified, the threshold setting has been taken as 10th/90th.

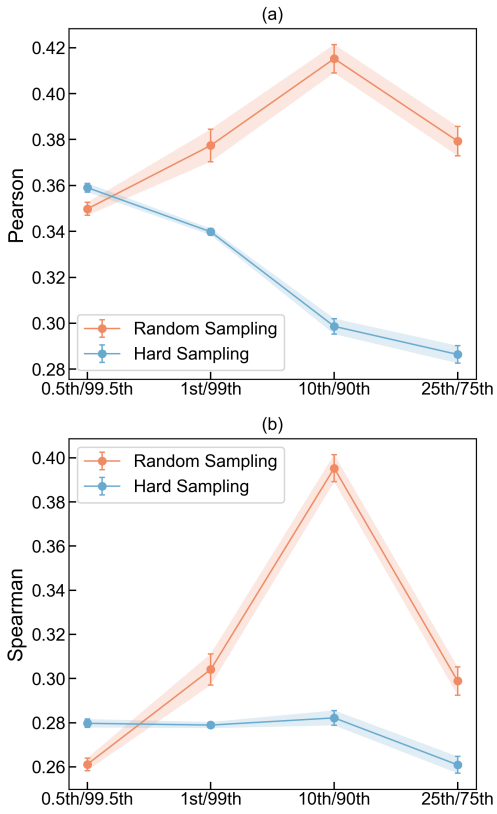

Figure 4: Performance variation with sampling methods and thresholds.

Triplet Sampling We implement two strategies, random sampling and hard sampling, to construct triplets of anchor-positive-negative paragraphs (s_A, s_A^+, s_A^-) from the training set. In random sampling, we select one positive instance s_A^+ and one negative instance s_A^- randomly from the respective positive pairs $(s_A \times \mathcal{S})^+$ and negative pairs $(s_A \times \mathcal{S})^-$ of the anchor s_A. In hard sampling, we utilize the pre-trained model $f_\theta(\cdot)$, which is later fine-tuned on these triplets, to embed paragraphs and calculate cosine similarities between the positive and negative pairs as $\textbf{cos}(f_\theta(s_A), f_\theta(s_A^{+/-}))$. For the positive instance, we choose the paragraph with the lowest similarity to the anchor from its positive pairs, that is,

$$s_A^+ = \underset{s}{\text{argmin}} \{\textbf{cos}(f_\theta(s_A),\ f_\theta(s)) \mid (s_A, s) \in (s_A \times S)^+\}$$

Conversely, for the negative instance, we select the paragraph with the highest similarity to the anchor from its negative pairs, that is,

$$s_A^- = \underset{s}{\text{argmax}} \{\textbf{cos}(f_\theta(s_A),\ f_\theta(s)) \mid (s_A, s) \in (s_A \times S)^-\}$$

To prevent the model from over-fitting, we ensure that each paragraph is used as an anchor only once, applying this rule across both random and hard sampling strategies. We also compare the two sampling ways in Figure 4 with respect to each positive-negative splitting thresholds. It's interesting to find that the random sampling has been better than hard sampling ever since the threshold is higher/lower than 0.5th/99.5th, we note that the case could be due to the noise in dataset, which makes the hard sampling failed to find more helpful instances.

Fine-tuning with Contrastive Learning We fine-tune several pretrained sentence embedding models on the C-HIS-PSY training set, using a triplet loss function,

$$L_{triplet}(\theta) = \sum_{s_A \in S} \textbf{max}\{\mathcal{D}^+ - \mathcal{D}^-, 0\}$$

where $\mathcal{D}^+$ denotes the distance between the positive pair, i.e. $\|f_\theta(s_A) - f_\theta(s_A^+)\|_2^2$, and $\mathcal{D}^-$ denotes the distance between the negative pair, i.e. $\|f_\theta(s_A) - f_\theta(s_A^-)\|_2^2$, α is a constant set to be 5, and θ stands for the pre-trained weights to be fine-tuned. This loss function aims to minimize the squared Euclidean norm between the anchor and positive, and maximize the squared Euclidean norm between the anchor and negative.

We construct triplets from the C-HIS-PSY validation set to validate the models during training, performing a hyperparameter sweep, to select the best-performing configuration, as shown in Table 1. The performance metrics of all models substantially improved after fine-tuning, as shown in Figure 1.

Table 1: Fine-tuning models' results over *validation* split. We show the best performing configuration selected over the validation split which was the final configuration used to report each models' test performance.

Framework	Base Model	If Specific to Classical Chinese	Batch Size	Warmup Epochs	Learning Rate	Pearson	Spearman
BERT	Bert-ancient-chinese	✔	32	3	1.0e-05	**.43**	**.42**
RoBERTa	Guwenbert-base	✔	32	2	2.0e-05	.30	.37
	Guwenbert-large	✔	16	1	2.0e-05	.29	.30
SBERT	Paraphrase-multilingual-MiniLM-L12-v2	✗	32	1	2.0e-05	.19	.19
MacBERT+CoSENT	text2vec-base-chinese	✗	32	2	2.0e-05	.34	.32
ERNIE+CoSENT	text2vec-base-chinese-paraphrase	✗	32	2	2.0e-05	.40	.40
LERT+CoSENT	text2vec-large-chinese	✗	16	2	2.0e-05	.36	.37

Table 2: Fine-tuning models' final performance under three methods of DDR, CCR, and Prompting.

Framework	Base Model	Semantic Textual Similarity (*Easy Task*)		Semantic Textual Similarity (*Hard Task*)		Questionnaire Item Classification	Psychological Measure	
		Pears.	Spear.	Pears.	Spear.	Accuracy	Pears.	Spear.
(a) DDR								
Word2Vec (CBOW)	/	$.02_{\pm.11}$	$.02_{\pm.10}$	$-.03_{\pm.02}$	$-.02_{\pm.01}$	$.80_{\pm.16}$	$.24_{\pm.07}$	$.25_{\pm.05}$
Word2Vec (Skip-gram)	/	$.08_{\pm.11}$	$.09_{\pm.11}$	$.02_{\pm.02}$	$.02_{\pm.01}$	$.87_{\pm.15}$	$.17_{\pm.09}$	$.19_{\pm.08}$
FastText (CBOW)	/	$.05_{\pm.11}$	$.04_{\pm.10}$	$-.01_{\pm.01}$	$.01_{\pm.01}$	$.90_{\pm.13}$	$.24_{\pm.06}$	$.26_{\pm.05}$
FastText (Skip-gram)	/	$.10_{\pm.10}$	$.11_{\pm.10}$	$.03_{\pm.02}$	$.04_{\pm.01}$	$.85_{\pm.16}$	$.18_{\pm.08}$	$.20_{\pm.07}$
GloVe	/	$.07_{\pm.10}$	$.09_{\pm.11}$	$.01_{\pm.02}$	$.01_{\pm.01}$	$.83_{\pm.15}$	$.16_{\pm.09}$	$.21_{\pm.06}$
(b) CCR								
BERT	Bert-ancient-chinese	$\mathbf{.53}_{\pm.07}$	$\mathbf{.55}_{\pm.07}$	$\mathbf{.42}_{\pm.01}$	$\mathbf{.43}_{\pm.01}$	$.93_{\pm.11}$	$\mathbf{.29}_{\pm.08}$	$\mathbf{.29}_{\pm.08}$
RoBERTa	Guwenbert-base	$.29_{\pm.07}$	$.46_{\pm.09}$	$.25_{\pm.01}$	$.40_{\pm.01}$	$.90_{\pm.11}$	$.15_{\pm.05}$	$.15_{\pm.09}$
RoBERTa	Guwenbert-large	$.41_{\pm.05}$	$.44_{\pm.07}$	$.28_{\pm.01}$	$.31_{\pm.01}$	$.83_{\pm.13}$	$.23_{\pm.01}$	$.21_{\pm.06}$
SBERT	Paraphrase-multilingual-MiniLM-L12-v2	$.20_{\pm.15}$	$.21_{\pm.14}$	$.18_{\pm.01}$	$.19_{\pm.01}$	$.82_{\pm.19}$	$.16_{\pm.01}$	$.14_{\pm.02}$
MacBERT+CoSENT	text2vec-base-chinese	$.41_{\pm.09}$	$.40_{\pm.09}$	$.32_{\pm.01}$	$.31_{\pm.01}$	$\mathbf{.95}_{\pm.08}$	$.15_{\pm.03}$	$.15_{\pm.02}$
ERNIE+CoSENT	text2vec-base-chinese-paraphrase	$.45_{\pm.09}$	$.45_{\pm.09}$	$.38_{\pm.01}$	$.37_{\pm.01}$	$.93_{\pm.11}$	$.10_{\pm.09}$	$.10_{\pm.09}$
LERT+CoSENT	text2vec-large-chinese	$.46_{\pm.12}$	$.47_{\pm.08}$	$.36_{\pm.01}$	$.38_{\pm.01}$	$.97_{\pm.07}$	$.21_{\pm.06}$	$.19_{\pm.06}$
(c) Prompting								
GPT	GPT-3.5	.08	.04	.26	.28	.63	$.05_{\pm.08}$	$.08_{\pm.10}$
GPT	GPT-4	.62	.52	.40	.30	.77	$.21_{\pm.09}$	$.23_{\pm.10}$

4 Evaluation and Results

In three tasks, we evaluated CCR (with sentence embedding models) and compared it with the standard DDR approach (with word embedding models) and the prompting method with LLMs. The results are shown in Table 2.

4.1 Semantic Understanding

Understanding of Historical Text: Semantic Textual Similarity For the CCR method, we embed whole paragraphs with sentence embedding models, and then calculate the cosine similarity between each pair of paragraphs. For the DDR method, we average the word vectors of all the words in the paragraph, and then calculate the cosine similarity between each pair of paragraphs. For the LLM-

prompting method, we craft a few-shot prompt (Figure 8) asking for a similarity score, ranging from 0 to 1, between each pair of paragraphs. As mentioned, similarities between the titles of each pair of paragraphs are used as the pseudo ground truth.

We construct paragraph pairs for evaluation from paragraphs in the C-HIS-PSY test set through two different sampling methods: random sampling (where a paragraph is randomly paired with any other paragraph to form pairs) and threshold sampling (where a paragraph is paired only with positive or negative samples filtered by a certain threshold), respectively. Due to the thresholded sampling, the constructed pairs are positive and negative samples for each other, with more significant differences between them; thus, we refer to it as the Easy Task. In contrast, the pairs formed through pure random sampling might contain samples that are ambiguous and unclear, making the test more challenging, which we call the Hard Task, as shown in Table 2.

Understanding of Questionnaire Item: Text Classification We convert several broadly accepted questionnaires from English into classical Chinese, including Collectivism, Individualism (Oyserman et al., 2002), Tightness and Looseness (Gelfand et al., 2006), by employing the Cross-lingual Questionnaire Conversion (CQC) approach described in Section 3.1.

For both the CCR and DDR methods, all the items from these questionnaires are embedded. Then we conduct 10-fold cross-validation, using Support Vector Machines (SVM) as the classifier, and text embeddings or averaged word vectors as features. For the prompting method, we craft a few-shot prompt (Figure 9) directly asking for classification.

4.2 Psychological Measure

For both CCR and DDR methods, we calculate the average cosine similarities between each paragraph in the test set and all the items in each questionnaire, representing the "loading score" of the paragraph on the questionnaire.

For the prompting method, we craft a few-shot prompt (Figure 10) king for a score, ranging from 0 to 1, to measure each paragraph with respect to the topic of each questionnaire. Items in each questionnaire are provided in the prompt.

We built a corresponding dictionary for each psychological construct. Average similarities between the title of each paragraph and all the terms in each dictionary are used as the pseudo ground truth.

5 Benchmarking: Traditionalism, Authority and Attitude toward Reform

Moral values and political orientations are closely intertwined (Federico et al., 2013; Kivikangas et al., 2021). For example, the attitude of individuals toward reforms, policy changes, and new legislation often reflects traditionalism, conservatism, and respect for authority (Hackenburg et al., 2023; Koleva et al., 2012). Those with stronger traditionalist views are more likely to identify with the existing social order and resist changes to the status quo (Osborne et al., 2023; Jost and Hunyady, 2005).

Officials' Attitudes toward Reform in the 11th Century Throughout Chinese history, there have been numerous instances of significant reforms, one of the most notable of which being the Wang Anshi's New Policies in the 11th century, which faced mixed reactions from officials. We draw upon a dataset manually compiled by Wang (2022), who annotated the attitudes of 137 major officials toward the reform.

Individual-level Measure of Traditionalism and Authority We extract writings of these officials documented in the *Complete Prose of the Song Dynasty*. Questionnaires of traditionalism (Samore et al., 2023) (Figure 12) and authority (Atari et al., 2023a) (Figure 11) are converted from English into classical Chinese, by employing the Cross-lingual Questionnaire Conversion (CQC), described in Section 3.1.

Employing the best-performing fine-tuned model, we use our CCR pipeline to measure the levels of traditionalism and attitudes toward authority expressed in their texts. For each individual official, results are aggregated by calculating the average score across all of their writings.

Results In support of the validity of our pipeline and based on our theoretical framework, we found a significant correlation (Figure 5) between officials' attitudes toward the reforms and the levels of traditionalism and authority measured through CCR. Authority and traditionalism both show a significant negative correlation with support for reform, with Spearman correlation coeffecients less than 0.4 and p-values less than 0.001. Officials

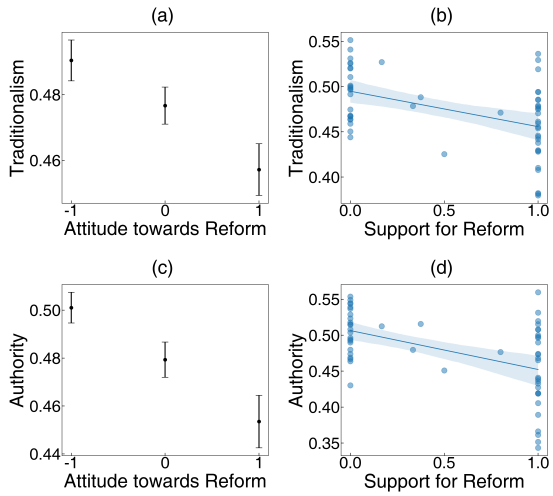

Figure 5: Correlation between Traditionalism / Authority and Officials' Attitudes toward Reforms. (a) and (c) present the average psychological measure scores with standard errors, utilizing an ordinal variable where -1 signifies opposition to the reform, 0 indicates a neutral or no explicit attitude, and 1 denotes support for the reform (Number of observations: 108). (b) and (d) depict the linear regression lines accompanied by 95% confidence intervals, employing a continuous variable that ranges from 0 to 1 to quantify officials' degree of support for the reform (Number of observations: 56).

with greater traditionalism and respect for existing authority are more likely to oppose reform.

Table 3: Spearman Correlation between CCR-based measure of moral values and actual attitude toward reform of officials.

	Support for Reform	Attitude toward Reform
Traditionalism	-0.441***	-0.279**
Authority	-0.472***	-0.310**

This finding supports the validity of CCR as a valid text-analytic pipeline to extract meaningful psychological information from classical Chinese corpora.

6 Discussion and Conclusion

Historical-psychological text analysis is a new line of research focused on extracting different aspects of psychology from historical corpora using state-of-the-art computational methods (Atari and Henrich, 2023). Here, we create a new pipeline, CCR, as a helpful tool for historical-psychological text analysis. Evaluating our model against word embedding models (e.g., DDR) and more recent LLMs (e.g., GPT4), we demonstrated that CCR performs better than these alternatives while keeping its high level of interpretability and flexibility. Classical Chinese is of great historical significance, and the proposed approach can be particularly helpful in testing new insights about the "dead minds" who lived centuries or even millennia prior. We hope our tool motivates future work at the intersections of psychology, quantitative history, and NLP.

7 Limitation

The judgment of moral values often carries subjectivity, and the patterns learned from the model carry the bias of pre-training data. Due to the severe lack of fine-grained data available for training in the fields of classical Chinese literature and historical texts, the method of indirect supervised learning we adopt may lead to the model learning some noise from the data, affecting the model's performance. Compiling more finely annotated datasets manually is our future work direction.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2023. Perils and opportunities in using large language models in psychological research.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023a. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188.

Mohammad Atari and Joseph Henrich. 2023. Historical psychology. *Current Directions in Psychological Science*, 32(2):176–183.

Mohammad Atari, Ali Omrani, and Morteza Dehghani. 2023b. Contextualized construct representation: Leveraging psychometric scales to advance theory-driven text analysis.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *ArXiv*, abs/2009.10053.

Nicolas Baumard, Lou Safra, Mauricio Martins, and Coralie Chevallier. 2024. Cognitive fossils: using cultural artifacts to reconstruct psychological changes throughout history. *Trends in Cognitive Sciences*, 28(2):172–186.

Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Ryan L Boyd and H Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

Virginia K Choi, Snehesh Shrestha, Xinyue Pan, and Michele J Gelfand. 2022. When danger strikes: A linguistic tool for tracking america's collective response to threats. *Proceedings of the National Academy of Sciences*, 119(4):e2113891119.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christopher M. Federico, Christopher R. Weber, Damla Ergun, and Corrie Hunt. 2013. Mapping the connections between politics and morality: The multiple sociopolitical orientations involved in moral intuition. *Political Psychology*, 34(4):589–610.

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior research methods*, 50:344–361.

Michele J Gelfand, Lisa H Nishii, and Jana L Raver. 2006. On the nature and importance of cultural tightness-looseness. *Journal of applied psychology*, 91(6):1225.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Kobi Hackenburg, William J Brady, and Manos Tsakiris. 2023. Mapping moral language on us presidential primary campaigns reveals rhetorical networks of political division and unity. *PNAS nexus*, page pgad189.

Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A. Lindquist. 2021. From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17(3):805–826.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

John T Jost and Orsolya Hunyady. 2005. Antecedents and consequences of system-justifying ideologies. *Current directions in psychological science*, 14(5):260–265.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511.

J Matias Kivikangas, Belén Fernández-Castilla, Simo Järvelä, Niklas Ravaja, and Jan-Erik Lönnqvist. 2021. Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological Bulletin*, 147(1):55.

Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. 2012. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of research in personality*, 46(2):184–194.

Zhou Liu, Hongsu Wang, and Peter K Bol. 2023. Automatic biographical information extraction from local gazetteers with bi-lstm-crf model and bert. *International Journal of Digital Humanities*, 4(1-3):195–212.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Michael Muthukrishna, Joseph Henrich, and Edward Slingerland. 2021. Psychology as a historical science. *Annual Review of Psychology*, 72(1):717–749.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Danny Osborne, Thomas H. Costello, John Duckitt, and Chris G. Sibley. 2023. The psychological causes and societal consequences of authoritarianism. *Nature Reviews Psychology*, 2(4):220–232.

Daphna Oyserman, Heather M. Coon, and Markus Kemmelmeier. 2002. Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin*, 128(1):3–72.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Fanchao Qi, Yanhui Yang, Jing Yi, Zhili Cheng, Zhiyuan Liu, and Maosong Sun. 2022. QuoteR: A benchmark of quote recommendation for writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 336–348, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Theodore Samore, Daniel M. T. Fessler, Adam Maxwell Sparks, Colin Holbrook, Lene Aarøe, Carmen Gloria Baeza, María Teresa Barbato, Pat Barclay, Renatas Berniūnas, Jorge Contreras-Garduño, Bernardo Costa-Neves, Maria del Pilar Grazioso, Pınar Elmas, Peter Fedor, Ana Maria Fernandez, Regina Fernández-Morales, Leonel Garcia-Marques, Paulina Giraldo-Perez, Pelin Gul, Fanny Habacht, Youssef Hasan, Earl John Hernandez, Tomasz Jarmakowski, Shanmukh Kamble, Tatsuya Kameda, Bia Kim, Tom R. Kupfer, Maho Kurita, Norman P. Li, Junsong Lu, Francesca R. Luberti, María Andrée Maegli, Marinés Mejia, Coby Morvinski, Aoi Naito, Alice Ng'ang'a, Angélica Nascimento de Oliveira, Daniel N. Posner, Pavol Prokop, Yaniv Shani, Walter Omar Paniagua Solorzano, Stefan Stieger, Angela Oktavia Suryani, Lynn K. L. Tan, Joshua M. Tybur, Hugo Viciana, Amandine Visine, Jin Wang, and Xiao-Tian Wang. 2023. Greater traditionalism predicts covid-19 precautionary behaviors across 27 societies. *Scientific Reports*, 13(1).

Indira Sen, Daniele Quercia, Marios Constantinides, Matteo Montecchi, Licia Capra, Sanja Scepanovic, and Renzo Bianchi. 2022. Depression at work: exploring depression in major us companies from online reviews. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–21.

Almog Simchon, William J Brady, and Jay J Van Bavel. 2022. Troll and divide: the language of online polarization. *PNAS nexus*, 1(1):pgac019.

Almog Simchon, Britt Hadar, and Michael Gilead. 2023. A computational text analysis investigation of the relation between personal and linguistic agency. *Communications Psychology*, 1(1):23.

Edward Slingerland. 2013. Body and mind in early china: An integrated humanities–science approach. *Journal of the American Academy of Religion*, 81(1):6–55.

Edward Slingerland, Ryan Nichols, Kristoffer Neilbo, and Carson Logan. 2017. The distant reading of religious texts: A "big data" approach to mind-body concepts in early china. *Journal of the American Academy of Religion*, 85(4):985–1016.

Daniel Swanson and Francis Tyers. 2022. Handling stress in finite-state morphological analyzers for Ancient Greek and Ancient Hebrew. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 108–113, Marseille, France. European Language Resources Association.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiahui Wang, Xuqin Zhang, Jiahuan Li, and Shujian Huang. 2023. Pre-trained model in Ancient-Chinese-to-Modern-Chinese machine translation. In *Proceedings of ALT2023: Ancient Language Translation Workshop*, pages 23–28, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Sze-Yuh Nina Wang and Yoel Inbar. 2021. Moral-language use by us political elites. *Psychological Science*, 32(1):14–26.

Yuhua Wang. 2022. Blood is thicker than water: Elite kinship networks and state building in imperial china. *American Political Science Review*, 116(3):896–910.

John Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20:529–544.

Mengyao Xu, Lingshu Hu, and Glen T Cameron. 2023. Tracking moral divergence with ddr in presidential debates over 60 years. *Journal of Computational Social Science*, 6(1):339–357.

Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Ying Zhong, Valentin Thouzeau, and Nicolas Baumard. 2023. The evolution of romantic love in chinese fiction in the very long run (618 - 2022): A quantitative approach. In *Workshop on Computational Humanities Research*.

Ke Zhou, Luca Maria Aiello, Sanja Scepanovic, Daniele Quercia, and Sara Konrath. 2021. The language of situational empathy. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–19.

A Historical Psychology Corpus Details

A.1 Distribution of Paragraph Lengths

To ensure the inclusion of sufficient semantic information, paragraphs containing fewer than 50 characters have been merged with the preceding paragraph of the article or chapter, wherever possible. To accommodate the token limitations of models such as BERT, paragraphs that exceed 500 characters have been divided into segments with fewer than 500 characters each, while maintaining the integrity of the original sentence structure as much as possible. The average length of paragraphs is 195 characters.

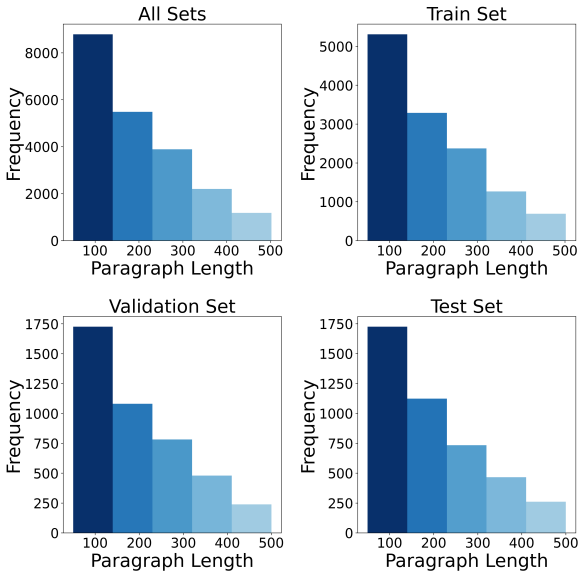

Figure 6: Distributions of paragraph lengths in different sets.

A.2 Distribution of Title Similarities

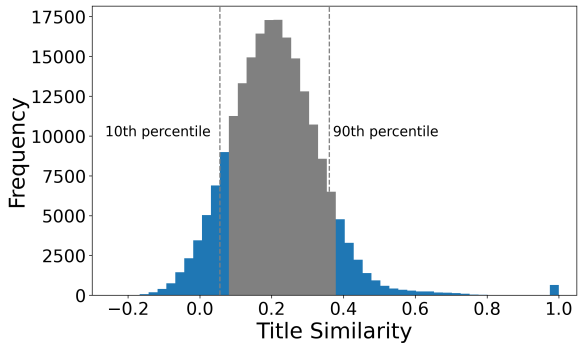

Figure 7: Distribution of title similarities with thresholds.

B Word Embedding Model Details

B.1 Corpus

B.2 Pre-processing

Before training the word vector model, we conducted word segmentation on the corpus, employing the pretrained tokenizer "COARSE_ELECTRA_SMALL_ZH" from HanLP (https://hanlp.hankcs.com/docs/api/hanlp/pretrained/tok.html).

After word segmentation, the corpus consists of 1.04 billion word tokens and an initial vocabulary containing 15.55 million unique words. By truncating the vocabulary at a minimum word count threshold of 10, the final vocabulary size is reduced to 1.27 million words.

B.3 Training Hyperparameters

We train our word vector models on the same corpus using various frameworks and architectures, such as Word2Vec (with CBOW and Skip-gram) (Mikolov et al., 2013), FastText (with CBOW and Skip-gram) (Bojanowski et al., 2017), and GloVe (Pennington et al., 2014). The hyperparameters are presented in Table 4.

Instructions

```
On a scale ranging from 0 to 1, how similar are the
following two paragraphs with respect to moral
values? Respond only with a score ranging from 0 to 1.
```

Examples

```
Here are two examples:

Input:
Paragraph 1: 蘇應喜，正安州人，年十八，母劉氏。道光乙酉，東街火延及
西街，民居殆盡。喜方在書院肄業，聞報奔回，不問物，惟尋母耗。不得，既而聞
火中哭聲，喜知是母，急入救。眾以火猛，入必死，挽之，喜哭曰：「天下豈有無
母之子哉？」奮身入救，死之。後灰燼中見喜覆母，母通身焦黑，而喜面如生。

Paragraph 2: 十歲喪父，事母克盡孝道。母病，藥餌無效，振乃焚香籲天，
刲股肉調羹以進，母病遂痊。事聞，縣令旌曰真孝格天。

Output:
0.74346477

Input:
Paragraph 1: 蘇應喜，正安州人，年十八，母劉氏。道光乙酉，東街火延及
西街，民居殆盡。喜方在書院肄業，聞報奔回，不問物，惟尋母耗。不得，既而聞
火中哭聲，喜知是母，急入救。眾以火猛，入必死，挽之，喜哭曰：「天下豈有無
母之子哉？」奮身入救，死之。後灰燼中見喜覆母，母通身焦黑，而喜面如生。

Paragraph 2: 東坡在儋耳，語其子過曰：「我決不為海外人。近日頗覺有還
中州氣象。」乃滌研焚香，寫平日所作八賦，當不脫誤一字以卜之。寫畢，大喜
曰：「吾歸無疑矣。」後數日，廉州之命至。八賦墨跡初歸梁師成，後入禁中。

Output:
0.043439843
```

Figure 8: Few-shot prompt for the semantic textual similarity task.

Instructions

```
There are four topics that are related to moral
values: "Collectivism", "Individualism", "Tightness",
and "Looseness". Respond with the topic that best fits
the following paragraph.
```

Examples

```
Here are two examples:

Input:
忠孝友悌，正己化人，矜孤恤寡，敬老懷幼。
Output: Collectivism

Input:
禮者禁於將然之前，而法者禁于已然之後。
Output: Tightness
```

Figure 9: Few-shot prompt for the questionnaire item classification task.

Instructions

```
The Collectivism questionnaire is as follows: 1.人人親其親，長其
長，而天下平。2.教以慈睦，而民貴有親。3.忠孝友悌，正己化人，矜孤恤寡，敬
老懷幼。4.君仁臣忠，父慈子孝，兄愛弟敬，夫和妻柔，姑慈婦聽，禮之至也。5.
事父母者莫善於順，宜兄弟者莫善於讓。故順，孝之實也；讓，友之本也。

The Individualism questionnaire is as follows: 1.餘凡事喜獨出己見，
不屑隨人是非。2.人生何必同？要在有所立。3.寧頸獨倡兮不與眾同。4.人須有
自信之能力，當從自己良心上認定是非，不可以眾人之是非為從違。5.凡事須先求
自己，臨場何必仰他人。

The Tightness questionnaire is as follows: 1.不設規矩，無以順人；不
切刑罰，無以息暴。2.法律政令者，吏民規矩繩墨也。3.峻法，所以凌過遊外私
也；嚴刑，所以遂令懲下也。4.刑一而正百，殺一而慎萬。5.以刑止刑，以殺止
殺。

The Looseness questionnaire is as follows: 1.人不擾，政已和。世慮
寡，山情多。2.我清靜則民自正。3.無為，無我，無欲，居下，清虛，自然。4.
思想自由，為凡百自由之母者。5.我無為，人自寧。

On a scale ranging from 0 to 1, what score would the
following text receive in these
four questionnaires respectively? Respond only with
four scores, ranging from 0 to 1, in the same order as
the questionnaires.
```

Examples

```
Here are two examples:

Input:
聶銖，字積道，子述之後，隱居盱水，不求聞達。以道誼相尚，朝野鉅公，聞人多
就。見焉，或勸其出，答曰：吾讀祖父書，樂堯舜道，平生之志願得矣。出處之
際，非吾所固必也。一舉不合，遂自隱。後辟薩，皆不就。所著有《昌江風土
記》。通六經，精子史，尤長於《春秋》發明奧旨，有功後學。老於巖穴，可謂博
雅君子。侔七世孫。性孝友，好學不于利祿，放情山水，有司以賢良薦，辭弗就。
Output:
0.23639056, 0.33721533, 0.12832133, 0.4139657

Input:
濟南西關有約承基者，年十五，父歿，祖母尚存，年八十矣。家貧，自知祖孫難以
存活，因傭於修造工程處，日得工錢三百文以養祖母。
Output:
0.37603232, 0.28392938, 0.1424624, 0.0230375
```

Figure 10: Few-shot prompt for the psychological measure task.

Table 4: Word Vector Model Training Hyperparameters and Evaluation Results

Framework	Architecture	Vector Size	Epoch	Window Size	Other Parameters
Word2Vec	CBOW	300	5	5	negative=5
	Skip-gram	300	5	5	negative=5
FastText	CBOW	300	5	5	negative=5, min_n=1, max_n=4
	Skip-gram	300	5	5	negative=5, min_n=1, max_n=4
GloVe		300	15	5	x_max=100, alpha=0.75

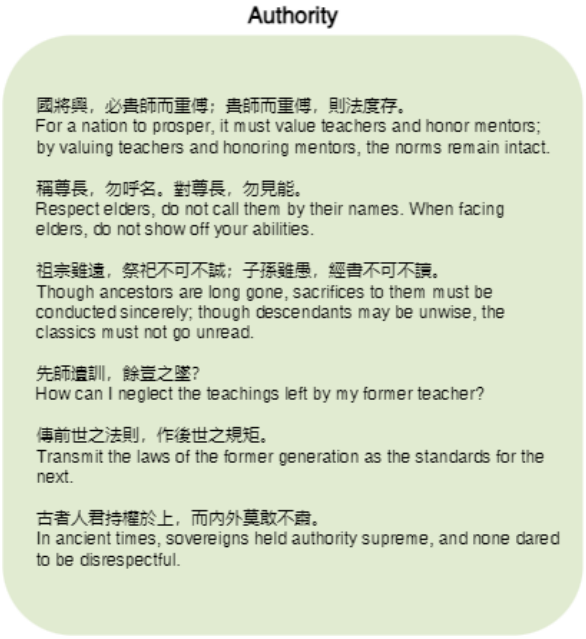

Figure 11: Questionnaire of Authority in classical Chinese.

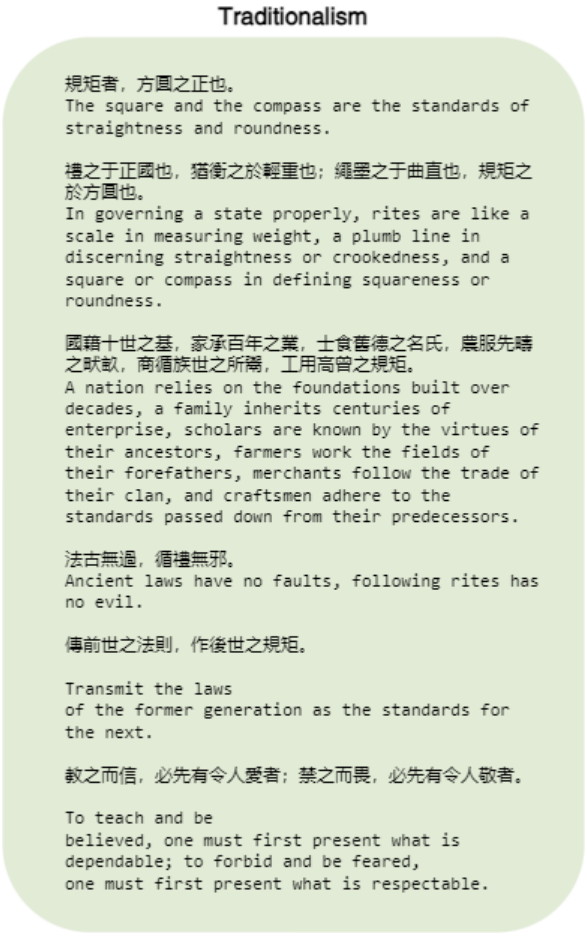

Figure 12: Questionnaire of Traditionalism in classical Chinese.