scientific data

OPEN

DATA DESCRIPTOR

ChineseEEG: A Chinese Linguistic Corpora EEG Dataset for Semantic Alignment and Neural Decoding

Xinyu Mou[1,10], Cuilin He[2,10], Liwei Tan[2,10], Junjie Yu[1], Huadong Liang[3], Jianyu Zhang[1], Yan Tian[2], Yu-Fang Yang[4], Ting Xu[5], Qing Wang[6], Miao Cao[7], Zijiao Chen[8], Chuan-Peng Hu[9], Xindi Wang[1], Quanying Liu[1] ✉ & Haiyan Wu[2] ✉

An Electroencephalography (EEG) dataset utilizing rich text stimuli can advance the understanding of how the brain encodes semantic information and contribute to semantic decoding in brain-computer interface (BCI). Addressing the scarcity of EEG datasets featuring Chinese linguistic stimuli, we present the ChineseEEG dataset, a high-density EEG dataset complemented by simultaneous eye-tracking recordings. This dataset was compiled while 10 participants silently read approximately 13 hours of Chinese text from two well-known novels. This dataset provides long-duration EEG recordings, along with pre-processed EEG sensor-level data and semantic embeddings of reading materials extracted by a pre-trained natural language processing (NLP) model. As a pilot EEG dataset derived from natural Chinese linguistic stimuli, ChineseEEG can significantly support research across neuroscience, NLP, and linguistics. It establishes a benchmark dataset for Chinese semantic decoding, aids in the development of BCIs, and facilitates the exploration of alignment between large language models and human cognitive processes. It can also aid research into the brain's mechanisms of language processing within the context of the Chinese natural language.

Background & Summary

The human brain's ability to rapidly comprehend linguistic information and generate corresponding linguistic expressions is an indicator of its complex processing capabilities[1]. When exposed to linguistic stimuli, the human brain encodes the semantic information through neural activities[2]. By analyzing such neural activities, we can uncover the encoding mechanisms of semantics in the brain[3]. A variety of neural signals, including EEG, Functional Magnetic Resonance Imaging (fMRI), Electrocorticography (ECoG) are employed in language-related tasks, from academic research like investigating language processing mechanisms in the brain to practical applications like language decoding in BCI[4–9]. Recently, a lot of studies on neurolinguistics utilized both traditional machine learning methods and modern deep learning methods in NLP to explore linguistic-related problems[10–16]. However, these data-driven methods rely heavily on massive and comprehensive datasets[17]. In the field of NLP, it is relatively easy to collect large amounts of natural language data. In contrast, acquiring a large volume of neural signals generated in response to natural language stimuli poses significant challenges. To utilize the strong ability of modern data-driven methods, it is important to scale neural datasets to commensurate the state-of-the-art NLP to encompass the wide range of language expressions encountered in daily life. Among all neuroimaging techniques, EEG holds great potential to meet this demand. EEG is non-invasive and cost-effective[18], which allows the creation of long-duration neural signal datasets enriched

[1]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China. [2]Centre for Cognitive and Brain Sciences, Department of Psychology, Faculty of Social Sciences, University of Macau, Taipa, Macau SAR, China. [3]AI Research Institute, iFLYTEK Co., LTD, Hefei, China. [4]Division of Experimental Psychology and Neuropsychology, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany. [5]Center for the Integrative Developmental Neuroscience, Child Mind Institute, New York, NY, USA. [6]Shanghai Mental Health Center, School of Medicine, Shanghai Jiao Tong University, 600 S. Wanping Rd., Shanghai, 200030, China. [7]Australian National Imaging Facility and Swinburne Neuroimaging Facility, Swinburne University of Technology, Victoria, Australia. [8]Centre for Cognitive and Cognition, Yong Loo Lin School of Medicine, National University of Singapore, Kent Ridge, Singapore. [9]School of Psychology, Nanjing Normal University, Nanjing, China. [10]These authors contributed equally: Xinyu Mou, Cuilin He, Liwei Tan. ✉e-mail: liuqy@sustech.edu.cn; haiyanwu@um.edu.mo

with semantic information. Meanwhile, EEG features high temporal resolution[19], which enables it to precisely capture the brain's rapid dynamic changes in the language processing process.

Despite the abundance of EEG datasets for natural visual stimuli (e.g., THINGS-EEG)[20–23], those for natural language stimuli remain scarce. Currently, only a few language-related EEG datasets exist, such as the ZuCo dataset[24]. However, the majority of these datasets are collected using stimuli from English language corpora. This leads to limited research on the neural representations of other languages like Chinese. The brain's processing mechanisms differ for various languages. For example, the brain exhibits specificity in response to Chinese compared to English[25]. Therefore, it is important to create an EEG dataset based on other language stimuli. Chinese, being distinct from English in both structure and semantics, provides an opportunity to expand our understanding of neural responses to linguistic stimuli. An EEG dataset stimulated by Chinese corpora can facilitate the investigation of cross-linguistic commonalities and variations in language processing in the brain, bringing new perspectives to our understanding of language processing mechanisms.

To address these gaps, we have collected an EEG dataset, named the "ChineseEEG" (Chinese Linguistic Corpora EEG Dataset)[26,27]. It contains high-density EEG data and simultaneous eye-tracking data recorded from 10 participants, each silently reading Chinese text for about 13 hours. The text materials are sourced from two well-known novels, *the Little Prince* and *Garnett Dream*, both in their Chinese versions. This dataset further comprises multiple versions of pre-processed EEG sensor-level data generated under different parameter settings, offering researchers a diverse range of selections. Additionally, we provide embeddings of the Chinese text materials encoded from BERT-base-chinese model, which is a pre-trained NLP model specifically used for Chinese[28], aiding researchers in exploring the alignment between text embeddings from NLP models and brain information representations in neural signals.

ChineseEEG[26,27] is a pilot EEG dataset specifically stimulated by Chinese text. It offers several advantages. Firstly, each participant was exposed to around 13 hours of diverse Chinese linguistic stimuli, encompassing a broad spectrum of semantic information. The extensive exposure is significant for studying the long-term neural dynamics of language processing in the brain. Secondly, we employed 128 channels of high-density EEG data, which offers superior spatial resolution for precise localization of brain regions involved in language processing. Besides, with a sampling rate of 1 kHz, it effectively captures the dynamics of neural representations during reading. Furthermore, the inclusion of the pre-processed EEG data and text embeddings is beneficial for scholars in both neuroscience and computer science domains who lack inter-disciplinary experience, enabling them to directly utilize well processed data from fields they may not familiar with.

ChineseEEG[26,27] can serve as a valuable resource in neuroscience, linguistics, and other related fields. EEG data generated from Chinese language stimuli will significantly support research within the Chinese context, aiding researchers in revealing the characteristics of brain signal representations under Chinese stimuli, and promoting the development of brain-to-text translation, semantic decoding and other practical applications tailored to Chinese context. The dataset can also bring diversity to languages used in related research, encouraging the exploration of similarities and differences in language processing stimulated by different languages. It can also aid in multi-linguistic alignment in NLP by aligning multi-lingual brain signals with natural languages[29,30]. Given the dataset's inclusion of widely used text materials like *the Little Prince* in multilingual neuroimaging research, ChineseEEG can be combined with prior datasets to extend its potential. For example, combining ChineseEEG with neural signal datasets for auditory language comprehension tasks under similar semantic stimuli[31,32] can help to uncover the neural mechanisms of language understanding in multi-modal perceptions. Besides, researchers can integrate semantically rich EEG data in ChineseEEG with other neuroimaging modalities, such as fMRI and MEG, in language comprehension tasks[33,34] to precisely uncover brain's spatio-temporal dynamics and thus enhance the understanding of neural mechanisms of language processing in the brain.

Methods

Participants and task overview.

We recruited 15 participants (18–29 years old, averaged 21.9 years old, and 9 males). 3 participants participated the pre-experimental test before the official experiment to ensure the rationality of the experimental procedure and the stability of the devices. In the official experiment, 2 participants withdrew halfway due to scheduling conflicts (After communicating with the experimenter, they decided to withdraw from the experiment). In total, data from only 10 participants were used (18–29 years old, averaged 22.7 years old, and 5 males). No participant reported neurological or psychiatric history. All participants are right-handed and have normal or corrected-to-normal vision. Each participant voluntarily enrolled in and signed the informed consent form before the experiment and got a coupon compensation of approximately 50 MOP (MOP is the official currency of the Macao Special Administrative Region of China) for each experimental run (25 runs in total). This study complied with the Declaration of Helsinki and was performed according to the ethics committee approval of the Institutional Review Board of the University of Macau (Approval No. BSERE20-APP011-ICI).

Experimental material.

The experimental materials consist of two novels, both in the genre of children's literature. The first is the Chinese translation of *the Little Prince* (http://www.xiaowangzi.org/index.html) and the second is *Garnett Dream* (https://www.feiku6.com/read/s3-langwangmeng/18242419.html), both sourced from the Internet. Using novels, especially children's literature provides several advantages for research, especially within a naturalistic paradigm. Firstly, given their extensive size, these novels offer vast and diverse linguistic content, encompassing the majority of frequently utilized Chinese characters and daily expressions. Besides, children's literature can create an engaging environment for participants, making them more focused and emotionally engaged in the experiment.

Each novel was used as the material for a single session in the experiment. Each session was divided into several runs. For *the Little Prince*, the preface was used as the material for the practice reading phase. The main

Session	Run	Chapter	Number of Chinese characters	Duration
LittlePrince		Preface	210	
	1	1–4	3,805	24min34s
	2	5–8	3,734	24min5s
	3	9–12	3,218	20min50s
	4	13–16	4,030	25min59s
	5	17–20	1,713	11min11s
	6	21–24	3,635	23min27s
	7	25–27	4,189	26min54s
GarnettDream	1	1	5,267	34min17s
	2	2	4,406	28min39s
	3	3	5,327	34min35s
	4	4	3,906	25min15s
	5	5	4,989	32min14s
	6	6	4,413	28min29s
	7	7	3,912	25min25s
	8	8	5,537	35min52s
	9	9	4,171	27min2s
	10	10	5,943	38min30s
	11	11	4,351	28min21s
	12	12	4,830	31min13s
	13	13	3,799	24min31s
	14	14	4,963	32min9s
	15	15	4,656	29min55s
	16	16	4,615	29min42s
	17	17	5,273	33min57s
	18	18	5,113	32min57s
	19	19	5,438	35min10s

Table 1. An overview of the experiment.

body of the novel was then used for seven runs in the formal reading phase. The first six runs each includes 4 chapters of the novel, while the seventh run includes the last 3 chapters. For *Garnett Dream*, the first 18 chapters were used for 18 runs in the formal reading stage, with each run including a complete chapter. Due to the loss of markers during the EEG collection process, run 18 of ses-GarnettDream of sub-07 is unusable. We requested this participant to re-complete the reading task using chapter 19 of *Garnett Dream*.

To properly present the text on the screen during the experiment, the content of each run was segmented into a series of units, with each unit containing no more than 10 Chinese characters. These segmented contents were saved in Excel (.xlsx) format for subsequent usage. During the experiment, three adjacent units from each run's content will be displayed on the screen in three separate lines, with the middle line highlighted for the participant to read. The relevant code has been uploaded to the GitHub repository. See Code availability section for detailed information.

The overview of experimental materials is shown in Table 1. In summary, a total of 115,233 characters (24,324 in *the Little Prince* and 90,909 in *Garnett Dream*), of which 2,985 characters are unique, are used as experimental stimuli in ChineseEEG dataset.

Experimental procedures. Participants were instructed to sit in an adjustable chair, with their eyes positioned approximately 67 cm away from the monitor (Dell, width: 54 cm, height: 30.375 cm, resolution: 1,920 × 1,080 pixels, vertical refresh rate: 60 Hz), see Fig. 1b. They were tasked with reading a novel and were required to keep their heads still and keep their gaze on the highlighted (red) Chinese characters moving across the screen, reading at a pace set by the program. Eye-tracking technique was utilized to confirm that participants followed the highlighted characters.

Each participant was required to complete a total of 1 practice reading phase and 2 formal reading sessions. *the Little Prince* session was divided into 7 experimental runs and *Garnett Dream* session was divided into 18 experimental runs. The schedule for the entire experiment is as follows: participants were required to finish all experimental runs over the span of 8 days. The total daily reading duration was set at approximately 1.5 hours to avoid fatigue. Specifically, the reading tasks for the first day comprised the practice reading phase and runs 1–4 of *the Little Prince* session. The tasks for the second day comprised runs 5–8 of *the Little Prince* session. From the third to the eighth day, each day's reading tasks comprised 3 runs of *Garnett Dream* session. While participants were afforded the flexibility to adjust their schedules in the experiment, they were required to complete all reading tasks within one month.

Each experimental run lasted approximately 30 minutes and was divided into two phases: the eye-tracker calibration phase and the reading phase.

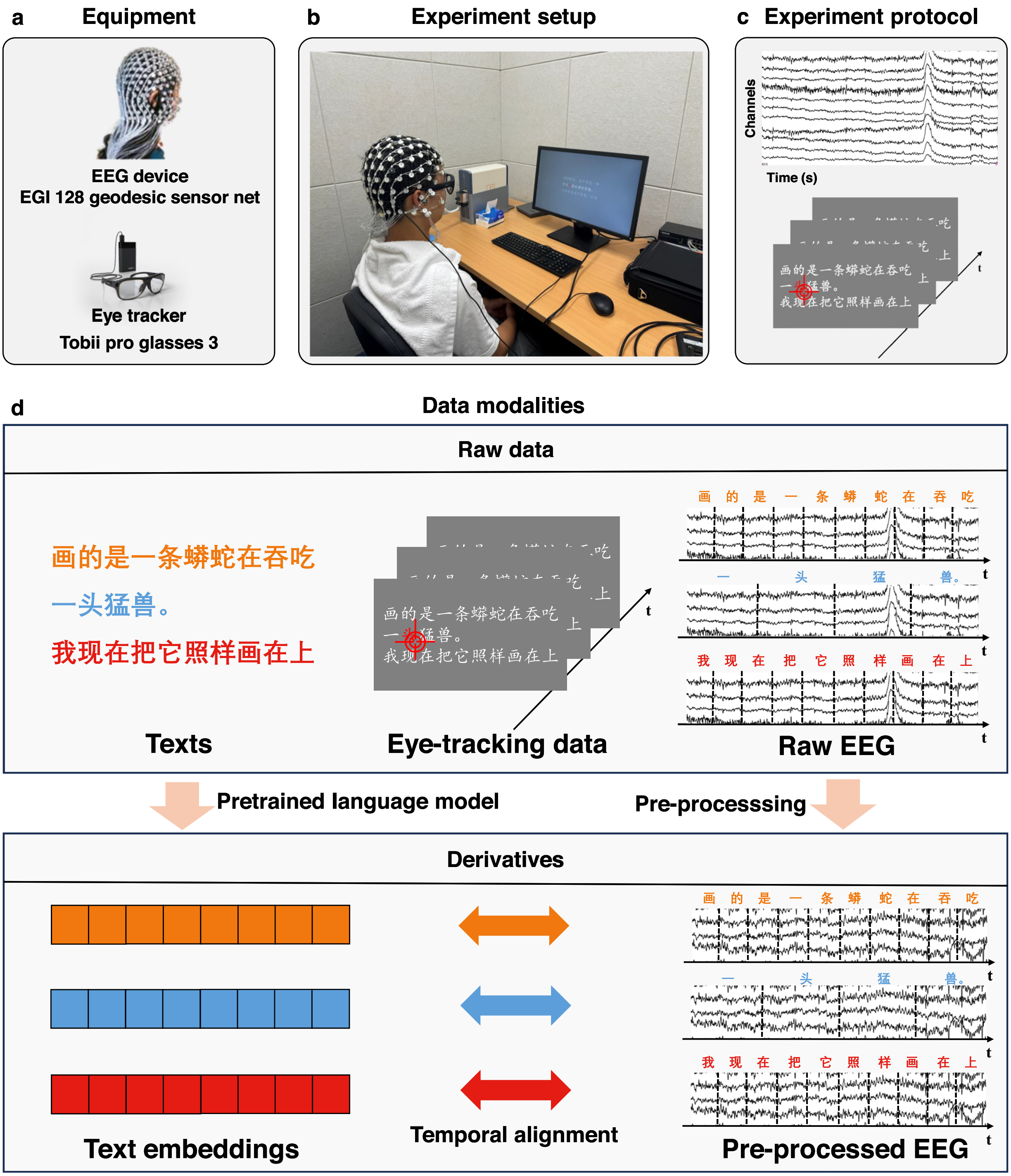


Fig. 1 Overview of the experiment and the modalities included in the dataset. (**a**) Equipment utilized in the experiment, including the EGI device for collecting EEG data and the Tobii Pro Glasses 3 eye-tracker for tracking eye movements. (**b**) The experiment setup. Participants were instructed to sit quietly approximately 67 cm from the screen and sequentially read the highlighted text. (**c**) The experimental protocol. Participants' 128-channel EEG signals and eye-tracking data were recorded while reading the highlighted text. (**d**) The data modalities in the dataset. The dataset comprises raw data such as the original textual stimuli, eye movement data, EEG data, and derivatives such as text embeddings from pre-trained NLP models and pre-processed EEG data.

Phase 1: Eye-tracker calibration phase. At the beginning of each run, participants were required to undergo an eye-tracker calibration process. Initially, the message "Hello! Please press the spacebar to start calibration" was displayed at the screen's center. Participants were instructed to keep their gaze at a fixation point, which sequentially appeared at the four corners and the center of the screen, each for 5 seconds. If the calibration failed, participants were prompted to start another calibration. Upon successful calibration, the message "Calibration successful! The page will automatically redirect in 5 seconds" was displayed at the center of the screen.

During the reading process, the accuracy of eye-tracking data can be influenced by several factors, including drift errors resulting from involuntary eye movements, as well as head movements and equipment positioning discrepancies. By performing calibrations at the beginning of each experimental run, these potential errors can be effectively mitigated, thereby ensuring the precision of the eye-tracking data.

Phase 2: Reading phase. After the calibration phase, participants were automatically directed to the reading phase. During the reading process, the screen initially displayed the serial number of the current chapter. Subsequently, the text appeared with three lines per page, ensuring each line contained no more than ten Chinese characters (excluding punctuation). On each page, the middle line was highlighted as the focal point, while the upper and lower lines were displayed with reduced intensity as the background. Each character in the middle line was sequentially highlighted with red color for 0.35 s, and participants were required to read the novel content following the highlighted cues. To facilitate a smooth reading experience, the text was designed to scroll automatically on the screen. Once participants finished reading the highlighted middle line, the text would scroll, moving the third line up to become the new middle line on the subsequent page.

The reading speed, which is slower than the typical speeds reported in previous studies[35], was deliberately chosen. This speed was selected based on feedback from the pre-experimental test to maintain participants' attention and minimize fatigue throughout the relatively long experimental run. The reading speed was fixed to enable character-level alignment between EEG segments and text. Additionally, fixed speed can also minimize the impact of external interference in the experiment and eliminate the impact of different reading speeds of different participants on subsequent analyses.

To ensure the accuracy of both EEG and eye-tracking data, participants were instructed to consistently focus on the highlighted text, while avoiding significant body movements to maintain a stable reading position. This protocol was strictly enforced to reduce any potential drifts and artifacts in the recordings.

After each run, participants were given sufficient time to rest. They were instructed to start the subsequent run only when they explicitly reported being ready to proceed. Adequate rest time can mitigate fatigue and enable participants to sustain their attention throughout the experiment, thus ensuring the quality of both EEG and eye-tracking data. The experimenter also evaluated each participant's performance and fatigue level through oral inquiries after each experimental run to ensure they could fully maintain their attention in subsequent runs. During each rest period, the experimenter replenished the saline solution on the electrodes of the EEG cap, which helped to maintain a low impedance, ensuring the collection of high-quality EEG data. Additionally, the experimenter checked the power status of the eye-tracker and replaced the batteries as necessary to ensure its continuous operation.

It should be noted that during the initial participation in the experiment, participants were required to complete a practice reading phase. The preface chapter of *the Little Prince* was selected as the reading material for this phase. All settings remained the same as those of the formal reading stage, to familiarize participants with the eye-tracker calibration process and the reading task.

The presentation of stimuli was managed using PsychoPy v2023.2.3[36], with the EGI PyNetstation v1.0.1 module facilitating the connection between PsychoPy and EGI Netstation. We also utilized g3pylib package to control our eye-tracker to follow the eye movement trajectories of the participants.

Data collection and analysis. This section shows the details of the data collection, pre-processing, and data analysis procedure. The modalities included in our dataset[26,27] are shown in Fig. 1d, including raw data and derivatives. Raw data contains the raw EEG data, eye-tracking data, raw text materials, and derivatives contain pre-processed EEG data and text embeddings generated by a pre-trained NLP model BERT-base-chinese.

EEG data collection. EEG data was acquired using an EGI 128-channel cap based on the GSN-HydroCel-128 montage with the Geodesic Sensor Net system (see Fig. 1a). The egi-pynetstation v1.0.1 package was used to control the EGI system. Before recording, the experimenter used a soft ruler to locate the position of the Cz electrode (i.e., the center of the brain) for each participant, ensuring the alignment of the electrodes in each experimental run. During recording, the sampling rate was 1 kHz. The impedance of each electrode was kept below 50 kΩ during the experiment. Setups and recording parameters are similar to our previous EEG dataset[37]. To precisely co-register EEG segments with individual characters during the experiment, we marked the EEG data with triggers (Table 2). The raw EEG data was exported to metafile format (.mff) files on the macOS system.

Eye-tracking data collection. Eye-tracking data was acquired using Tobii Pro Glasses 3. The device features 16 illuminators and 4 eye cameras integrated into scratch-resistant lenses, along with a wide-angle scene camera, allowing for a comprehensive capture of participant behavior and environmental context (see Fig. 1a). Due to the extensive duration of our experiments, the requirement for a lightweight eye-tracker was prioritized. The Tobii Pro Glass 3 fulfilled this criterion. Tobii Pro Glass 3 has a maximum sampling rate of 100 Hz. Given the relatively slow reading speed in our experiment, a sampling rate of 100 Hz is adequate for capturing the eye movement trajectories of the participants and assessing whether they were fixating on highlighted text at specific moments. More information about Tobii Pro Glass 3 can be found on the official website (https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3). We utilized the package g3pylib to control the glasses. The raw data was exported to .rar files.

EEG data pre-processing. To retain maximum amount of valid information in the data, we performed minimal pre-processing on the data, allowing researchers to further process the data according to their specific research

Trigger	Description
EYES	Start of eye-tracker recording
EYEE	End of eye-tracker recording
CALS	Start of the calibration stage before reading
CALE	End of the calibration stage
BEGN	Start of EEG data collection by the EGI device
STOP	Stop collecting EEG data
CHxx	Start of each chapter, where xx is the chapter number (e.g., the first chapter is CH01)
ROWS	Start of a new line of text
ROWE	End of a line
PRES	Start of the preface reading phase
PREE	End of the preface reading phase

Table 2. EEG triggers.

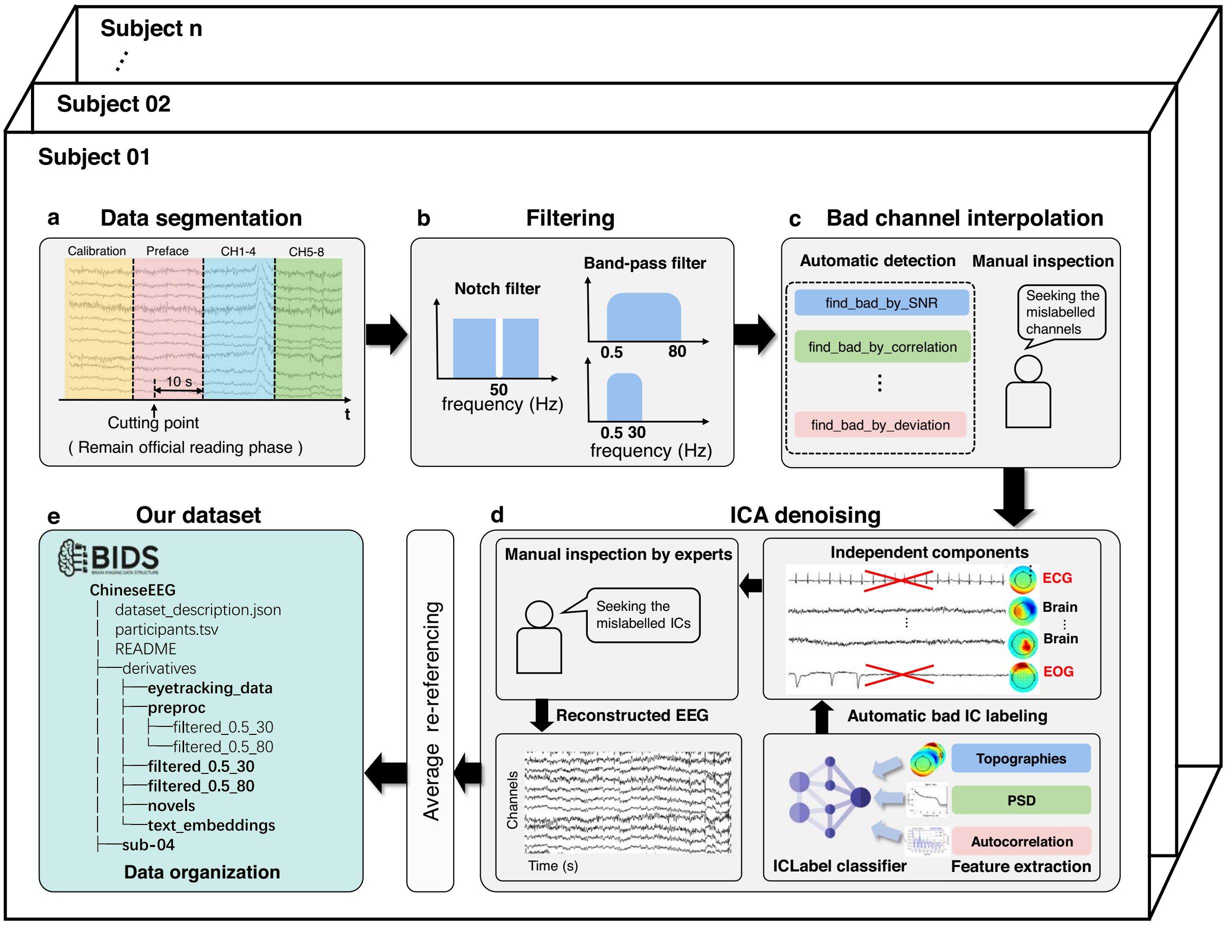


Fig. 2 EEG pre-processing pipeline. (**a**) Data segmentation: Data is segmented based on markers, retaining only the data from the formal reading phase. (**b**) Band-pass filtering: Two versions of filtered data are provided, with band-pass ranges of 0.5–30 Hz and 0.5–80 Hz respectively. (**c**) Bad channel interpolation: Our bad channel detection includes automatic detection implemented with the pyprep package and manual checking. For interpolation, the spherical spline interpolation implemented in MNE is utilized. (**d**) ICA denoising: In this part, the automatic labeling method in mne-iclabel package is utilized followed by a manual checking to remove noisy independent components such as eye movements and heartbeats. (**e**) Dataset organization: Our dataset is organized in the BIDS[41,42] format. The detailed file structure is shown in Fig. 3.

needs. The pre-processing pipeline is shown in Fig. 2. These pre-processing steps include data segmentation, downsampling, powerline filtering, band-pass filtering, bad channel interpolation, independent component analysis (ICA), and re-referencing. The MNE v1.6.0[38] package was utilized to implement all pre-processing steps.

During the data segmentation phase, we only retained data from the formal reading phase of the experiment. Based on the event markers during the data collection phase, we segmented the data, removing sections irrelevant to the formal experiment such as calibration and preface reading. To minimize the impact of subsequent filtering steps on the beginning and end of the signal, an additional 10 seconds of data was retained before the start of the formal reading phase. Subsequently, the signal was downsampled to 256 Hz. This specific sampling rate ensures effective capture of information related to language comprehension while reducing the burden of subsequent data processing and storage. Additionally, it aligns with the principle of minimal pre-processing, leaving necessary room for researchers to conduct personalized pre-processing based on their needs.

Following downsampling, a 50 Hz notch filter was applied to remove the powerline noise from the signal. Next, we performed band-pass overlap-add FIR filter on the signal to eliminate the low-frequency direct current components and high-frequency noise. Here, two versions of filtered data were offered. The first one has a filter band of 0.5–80 Hz and the second one has a filter band of 0.5–30 Hz. Researchers can choose the appropriate version based on their specific needs. After filtering, we performed an interpolation of bad channels. The bad channels were selected automatically using a Python-implemented EEG pre-processing package pyprep v0.4.3[39]. After automatic detection, we manually checked to avoid mislabeling or errors before interpolation. The spherical spline interpolation in the MNE package was utilized in this process.

Independent Component Analysis (ICA) was then applied to the data, utilizing the infomax algorithm available in the MNE package. The number of independent components was set to 20, ensuring that they contain the majority of information while not being so numerous to increase the burden of manual processing. Additionally, we set the random seed of the ICA algorithm to 97 to ensure the reproducibility of the ICA results. An automatic method was used to inspect and label components. It was implemented using mne-iclabel v0.5.1[40], which is a Python-implemented package for automatic independent component labeling. By manually inspecting the independent components after automatic labeling, we excluded obvious noise components such as Electrooculography (EOG) and Electrocardiogram (ECG). Finally, the data was re-referenced using the average method.

The process of manually identifying bad channels and excluding independent components during the ICA step can be conducted through annotations in a Graphical User Interface (GUI), making the annotation process quicker and more user-friendly.

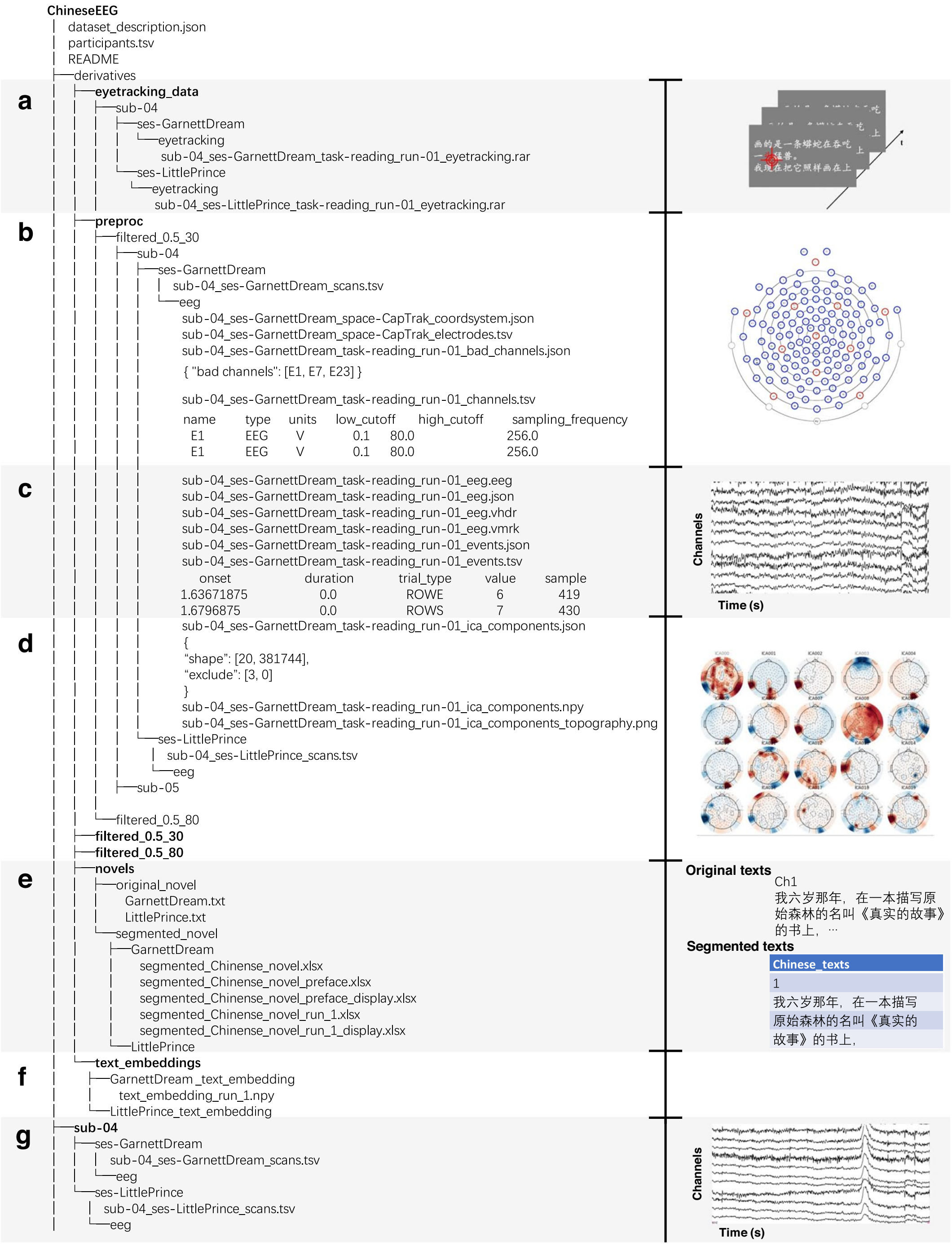
Data Records

The full dataset is publicly accessible via the ChineseNeuro Symphony community (CHNNeuro) in the Science Data Bank (ScienceDB) platform (https://doi.org/10.57760/sciencedb.CHNNeuro.00007)[26] or via the Openneuro platform (https://doi.org/10.18112/openneuro.ds004952.v1.2.0)[27]. Public data is distributed under the the Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/publicdomain/zero/1.0).

Data organization. The dataset[26,27] is organized following the EEG-BIDS[41,42] specification, which is an extension to the brain imaging data structure for EEG. The overview directory tree of our dataset is shown in Fig. 3. The dataset contains some regular BIDS files, 10 participants' data folders, and a *derivatives* folder. The stand-alone files offer an overview of the dataset: i) *dataset_description.json* is a JSON file depicting the information of the dataset, such as the name, dataset type and authors; ii) *participants.tsv* contains participants' information, such as age, sex, and handedness; iii) *participants.json* describes the column attributes in *participants.tsv*; iv) *README.md* contains a detailed introduction of the dataset.

Each participant's folder contains two folders named *ses-LittlePrince* and *ses-GarnettDream*, which store the data of this participant reading two novels, respectively. Each of the two folders contains a folder *eeg* and one file *sub-xx_scans.tsv*. The tsv file contains information about the scanning time of each file. The *eeg* folder contains the source raw EEG data of several runs, channels, and marker events files. Each run includes an *eeg.json* file, which encompasses detailed information for that run, such as the sampling rate and the number of channels. Events are stored in *events.tsv* with onset and event ID. The EEG data is converted from raw metafile format (*.mff* file) to BrainVision format (*.vhdr*,*.vmrk* and*.eeg* files) since EEG-BIDS[41,42] is not officially compatible with *.mff* format. All data is formatted to EEG-BIDS[41,42] using the mne-bids v0.14[42,43] package in Python.

The *derivatives* folder contains six folders: *eyetracking_data*, *filtered_0.5_80*, *filtered_0.5_30*, *preproc*, *novels*, and *text_embeddings*. The *eyetracking_data* folder contains all the eye-tracking data. Each eye-tracking data is formatted in a *.rar* file with eye moving trajectories and other parameters like sampling rate saved in different files. The *filtered_0.5_80* folder and *filtered_0.5_30* folder contain data that has been processed up to the pre-processing step of 0.5–80 Hz and 0.5–30 Hz band-pass filtering respectively. This data is suitable for researchers who have specific requirements and want to perform customized processing on subsequent pre-processing steps like ICA and re-referencing. The *preproc* folder contains minimally pre-processed EEG data that is processed using the whole pre-processing pipeline. It includes four additional types of files compared to the participants' raw data folders in the root directory: i) *bad_channels.json* contains bad channels marked during bad channel rejection phase. ii) *ica_components.npy* stores the values of all independent components in the ICA phase. iii) *ica_components.json* includes the independent components excluded in ICA (the ICA random seed is fixed, allowing for reproducible results). iv) *ica_components_topography.png* is a picture of the topographic maps of all independent components, where the excluded components are labeled in grey. The *novels* folder contains the original and segmented text stimuli materials. The original novels are saved in .txt format and the segmented novels corresponding to each experimental run are saved in Excel (.xlsx) files. The *text_embeddings* folder contains embeddings of the two novels. The embeddings corresponding to each experimental run are stored in NumPy (.npy) files.



```
ChineseEEG
│ dataset_description.json
│ participants.tsv
│ README
├─derivatives
│  ├─eyetracking_data
│  │  ├─sub-04
│  │  │  ├─ses-GarnettDream
│  │  │  │  └─eyetracking
│  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_eyetracking.rar
│  │  │  └─ses-LittlePrince
│  │  │     └─eyetracking
│  │  │        sub-04_ses-LittlePrince_task-reading_run-01_eyetracking.rar
│  ├─preproc
│  │  ├─filtered_0.5_30
│  │  │  ├─sub-04
│  │  │  │  ├─ses-GarnettDream
│  │  │  │  │  │ sub-04_ses-GarnettDream_scans.tsv
│  │  │  │  │  └─eeg
│  │  │  │  │     sub-04_ses-GarnettDream_space-CapTrak_coordsystem.json
│  │  │  │  │     sub-04_ses-GarnettDream_space-CapTrak_electrodes.tsv
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_bad_channels.json
│  │  │  │  │     { "bad channels": [E1, E7, E23] }
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_channels.tsv
│  │  │  │  │      name   type   units   low_cutoff   high_cutoff   sampling_frequency
│  │  │  │  │        E1   EEG    V          0.1      80.0          256.0
│  │  │  │  │        E1   EEG    V          0.1      80.0          256.0
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_eeg.eeg
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_eeg.json
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_eeg.vhdr
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_eeg.vmrk
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_events.json
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_events.tsv
│  │  │  │  │        onset          duration      trial_type     value    sample
│  │  │  │  │     1.63671875           0.0          ROWE          6        419
│  │  │  │  │     1.6796875            0.0          ROWS          7        430
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_ica_components.json
│  │  │  │  │     {
│  │  │  │  │      "shape": [20, 381744],
│  │  │  │  │      "exclude": [3, 0]
│  │  │  │  │     }
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_ica_components.npy
│  │  │  │  │     sub-04_ses-GarnettDream_task-reading_run-01_ica_components_topography.png
│  │  │  │  └─ses-LittlePrince
│  │  │  │     │ sub-04_ses-LittlePrince_scans.tsv
│  │  │  │     └─eeg
│  │  │  ├─sub-05
│  │  │
│  │  └─filtered_0.5_80
│  ├─filtered_0.5_30
│  ├─filtered_0.5_80
│  ├─novels
│  │  ├─original_novel
│  │  │    GarnettDream.txt
│  │  │    LittlePrince.txt
│  │  └─segmented_novel
│  │     ├─GarnettDream
│  │     │    segmented_Chinense_novel.xlsx
│  │     │    segmented_Chinense_novel_preface.xlsx
│  │     │    segmented_Chinense_novel_preface_display.xlsx
│  │     │    segmented_Chinense_novel_run_1.xlsx
│  │     │    segmented_Chinense_novel_run_1_display.xlsx
│  │     └─LittlePrince
│  └─text_embeddings
│     ├─GarnettDream _text_embedding
│     │    text_embedding_run_1.npy
│     └─LittlePrince_text_embedding
├─sub-04
│  ├─ses-GarnettDream
│  │  │ sub-04_ses-GarnettDream_scans.tsv
│  │  └─eeg
│  └─ses-LittlePrince
│     │ sub-04_ses-LittlePrince_scans.tsv
│     └─eeg
```

Original texts

Ch1
我六岁那年，在一本描写原始森林的名叫《真实的故事》的书上，…

Segmented texts

Chinese_texts
1
我六岁那年，在一本描写
原始森林的名叫《真实的
故事》的书上，

Fig. 3 File structure of the dataset. (**a**) Eye-tracking data: Each experimental run is associated with a .rar file that contains eye-tracking data. (**b**) Electrode information files: These include detailed information of electrodes such as the location, type, and sampling rate, as well as information on any channels marked as bad during pre-processing. (**c**) EEG data and event-related files: Including EEG data in BrainVision format and event files that record marker information. (**d**) ICA-related files: Containing independent components in numpy format, records of removed components during pre-processing, and topographic maps of the components. (**e**) Text materials: Containing original and segmented text. (**f**) Text embedding files: Each file corresponds to an experimental run and is stored in .npy format. (**g**) Raw EEG data.

Technical Validation

Classic sensor-level EEG analysis. The EEG data in the dataset[26,27] can be used to do classic time-frequency analysis. In this section, pre-processed EEG data was used to extract neural oscillations in different frequency bands. Specifically, we targeted the segment corresponding to the sentence "Draw me a sheep" in *the Little Prince* from the 0.5–80 Hz filtered pre-processed data of sub-07. The analysis was exclusively focused on the C3

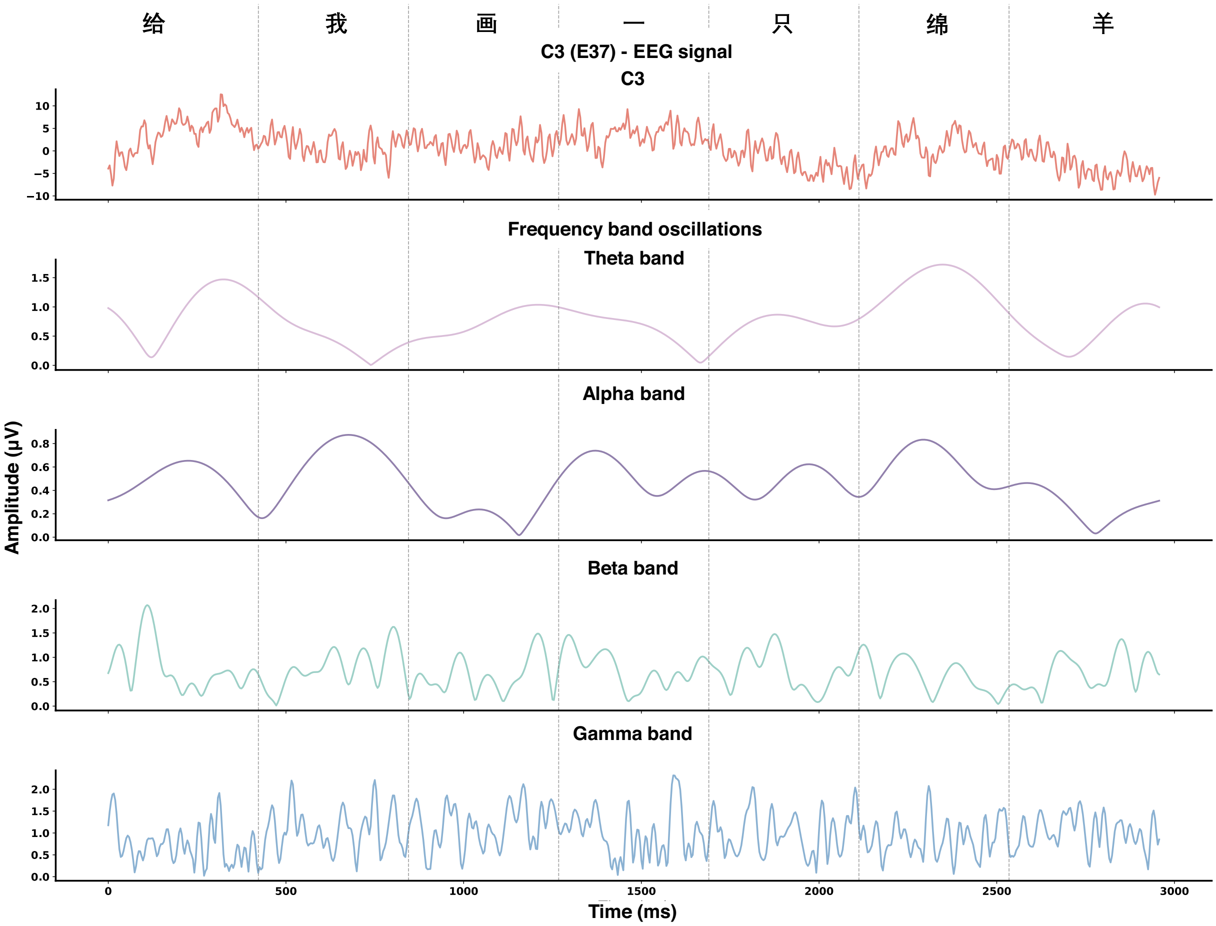


Fig. 4 EEG time course and the neural oscillations under different frequency bands (i.e., Theta, Alpha, Beta, and Gamma) corresponding to the Chinese sentence meaning "Draw me a sheep". The pre-processed EEG data using 0.5–80 Hz band-pass filter from ses-LittlePrince of sub-07 was used in the analysis. We illustrated the EEG signals from electrode C3, which locates at a language processing related area overlying the temporal lobe.

electrode to investigate the neural activities at the scalp location overlying the temporal lobe, which is a language processing related area.

To dissect the frequency components inherent in the C3 electrode's signal, we applied the Fast Fourier Transform (FFT) algorithm to the data. This mathematical technique transforms the time-domain signal into the frequency domain, revealing the spectrum of frequencies present in the neural recordings. We defined frequency bands of interest—Theta (4–8 Hz), Alpha (8–12 Hz), Beta (12–30 Hz), and Gamma (30–100 Hz)—to categorize the neural oscillations according to their respective frequency ranges.

For each frequency band, we separated the components from the FFT results and conducted an inverse FFT to retrieve the time-domain signal representing the band's oscillatory activity. This step allows for the quantitative analysis of the amplitude of oscillations within each frequency band, offering insights into the neurophysiological activity in these specific ranges. The results of different frequency bands are shown in Fig. 4.

EEG source reconstruction. Apart from the sensor level analysis, the EEG data allows for conducting source localization. Here, three segments of the data were utilized as an example to perform the source-level analysis using the MNE package. The fsaverage MRI template (https://surfer.nmr.mgh.harvard.edu/fswiki/FsAverage)[44] in MNE package was utilized to complete the surface reconstruction process. A 3-layer Boundary Element Method (BEM) model with 15360 triangles and conductivities of 0.3 S/m, 0.006 S/m, and 0.3 S/m for the brain, skull, and scalp compartments respectively was created. Source spaces consisted of 10242 sources per hemisphere. Three segments of the pre-processed EEG data with a band-pass frequency band of 0.5–80 Hz corresponding to one line displayed in the experiment were used to calculate the inverse solution. Inverse solutions were calculated using dynamic Statistical Parametric Maps (dSPM). The method was selected because it is widely used by researchers and is representative of currently used methods[45]. We offer the code of source reconstruction in our GitHub repository. See Code availability section for detailed information.

The visualization of the source activities is shown in Fig. 5b. Results for the left and right hemispheres are presented separately. The moments of peak activation in the left and right brain regions are chosen for visualization. The source localization results for the first segment reveal a dispersed activation area, encompassing the anterior temporal lobe and temporo-parietal region, which are associated with language comprehension and

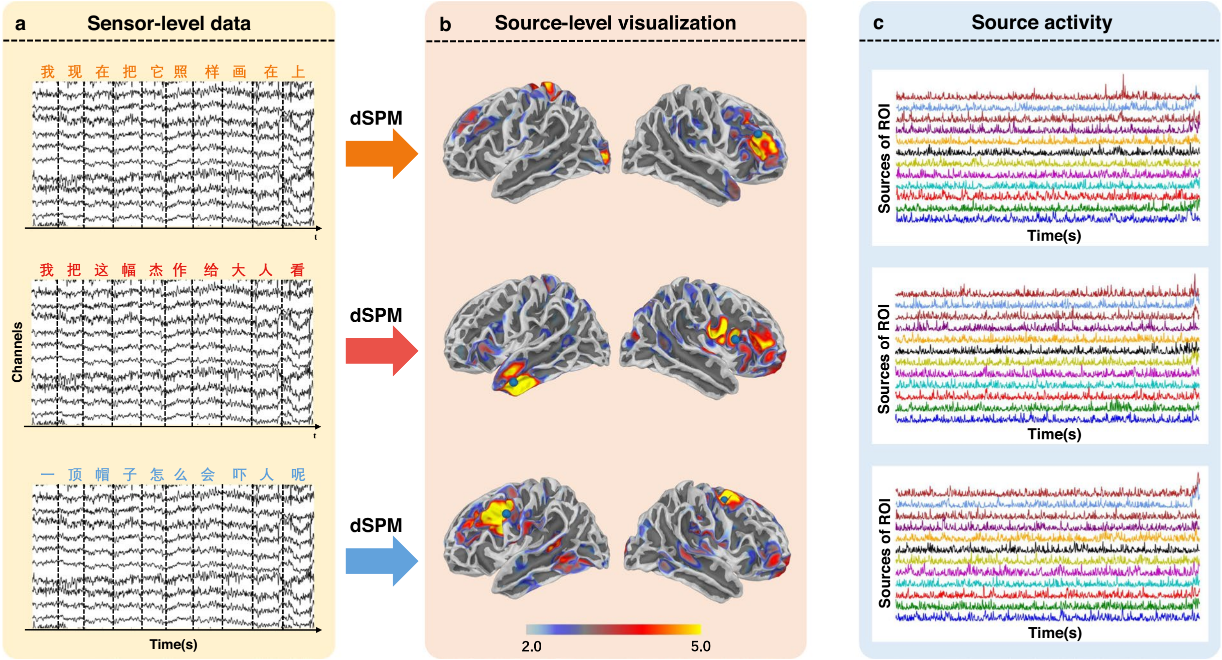


Fig. 5 EEG source localization analysis. (**a**) EEG sensor-level data: Three segments of pre-processed EEG data using 0.5–80 Hz band-pass filter were selected for analysis, accompanied by the corresponding text segments shown above the EEG segments. (**b**) Visualization of brain activation after source analysis: The dSPM method was utilized to solve the inverse problem. Results for the left and right hemispheres are presented separately. The moments of peak activation in the left and right brain regions are chosen for visualization. (**c**) Plots of source activity over time: Each plot contains the activities of 12 sources in the region with the strongest activity.

primary processing[46]. The results of the second segment exhibit more focused activation, particularly near the left middle temporal gyrus, an area (encompassing Wernicke's area) intimately related to language comprehension[47]. The activation areas for the third segment are localized in the left temporal and frontal lobes, potentially representing high-level stages of language processing, including sentence construction, semantic processing, and language expression[48]. Fig. 5c presents plots of source activities over time, derived from 12 sources in the corresponding region with strongest activities. The first two curves in each plot correspond to sources in the left and right hemispheres that reach maximum peak values.

Text embeddings with pre-trained language model.

To assist researchers in efficiently exploring the alignment between EEG and text representations, as well as in text decoding based on EEG, this study provides embeddings of two novels calculated using a pre-trained language model, accompanied by the code to compute these embeddings. This work employed Google's pre-trained language model BERT-base-Chinese[28]. This model, pre-trained on Chinese corpora, effectively encodes Chinese semantic features. Given that Chinese characters are the smallest unit of composition in Chinese writing and cannot be further decomposed[49], bert-base-chinese adopts a character-based tokenization approach, treating each Chinese character as a token for embedding. During the experimental procedure, each displayed line of text contains n Chinese characters. The BERT-base-Chinese model processes these n Chinese characters, yielding an embedding of size $(n, 768)$, where n represents the number of Chinese characters, and *768* the dimensionality of the embedding. To ensure displayed lines of varying length to have embeddings of the same shape, the first dimension of the embeddings is averaged to standardize the embedding size to *(1, 768)* for each instance. This processing procedure was implemented using the Hugging Face Transformers v4.36.2[50] package.

Temporal alignment between EEG, text sequences, and eye-tracking data.

This section provides a comprehensive explanation on how to align the EEG data with its corresponding text content and eye-tracking data in the temporal domain.

To facilitate semantic decoding, it is necessary to align specific text with its corresponding EEG segment in the temporal domain. During the marking process when collecting the data, the start and end of each line of the stimuli were annotated, thereby enabling the alignment of each text line with a corresponding segment of EEG data. Given the consistent highlighting duration for each character, the EEG segment can be equally divided to match the corresponding character. In the GitHub repository, we offer the script to align the EEG segments to their corresponding text and text embeddings.

The recorded eye-tracking data can be aligned with EEG data to verify whether participants were focusing on the text as expected. The eye-tracking data captures both the scene viewed by the participants at each moment and the coordinates of the gaze points. During each experimental run, the marker "EYES" and "EYEE" were inserted into the EEG recordings when the eye-tracker was activated and deactivated. These markers enable precise alignment between the eye-tracking data and the EEG recordings. Once the alignment is complete, markers in the EEG recordings enable the extraction of specific eye-tracking data segments corresponding to particular EEG segments. These eye-tracking data segments can be used to check whether the eye fixation locations align with the anticipated positions on the screen, thus reflecting the quality of the EEG data.

Usage Notes

The code for the experiment and data analysis has been uploaded to GitHub to facilitate sharing and utilization, which is accessible at https://github.com/ncclabsustech/Chinese_reading_task_eeg_processing.

The code repository contains four main modules, each including scripts desired to reproduce the experiment and data analysis procedures. The script *cut_chinese_novel.py* in the *novel_segmentation_and_text_embeddings* folder contains the code to prepare the stimulation materials from source materials. The script *play_novel.py* in the *experiment* module contains code for the experiment, including text stimuli presentation and control of the EGI device and Tobii Pro Glasses 3 eye-tracker. The script *preprocessing.py* in *data_preprocessing_and_alignment* module contains the main part of the code to apply pre-processing on EEG data. The script *align_eeg_with_sentence.py* in the same module contains code to align the EEG segments with corresponding text contents and text embeddings. The *docker* module contains the Docker image required for deploying and running the code, as well as tutorials on how to use Docker for environment deployment.

The code for EEG data pre-processing is highly configurable, permitting flexible adjustments of various pre-processing parameters, such as data segmentation range, downsampling rate, filtering range, and choice of ICA algorithm, thereby ensuring convenience and efficiency. Researchers can modify and optimize this code according to their specific requirements.

Before using our ChineseEEG dataset[26,27], we encourage all users to check the *README.md* and the updated information in the GitHub repository.

Code availability

The code for all modules is openly available on GitHub (https://github.com/ncclabsustech/Chinese_reading_task_eeg_processing). All scripts were developed in Python 3.10[51]. Package openpyxl v3.1.2 was utilized to export segmented text in Excel (.xlsx) files, and egi-pynetstation v1.0.1, g3pylib v0.1.1, psychopy v2023.2.3[36] were used to implement the scripts for EGI device control, Tobii eye-tracker control, stimuli presentation respectively. In the data pre-processing scripts, MNE v1.6.0[38], pybv v0.7.5[52], pyprep v0.4.3[39], mne-iclabel v0.5.1[40] were used to implement the pre-processing pipeline, while mne-bids v0.14[42,43] was used to organize the data into BIDS[41,42] format. The text embeddings were calculated using Hugging Face transformers v4.36.2[50]. For more details about code usage, please refer to the GitHub repository.

Received: 11 February 2024; Accepted: 21 May 2024;
Published online: 29 May 2024

References

1. Friederici, A. D. The brain basis of language processing: From structure to function. *Physiological Reviews* **91**, 1357–1392 (2011).
2. Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A. & Rogers, T. T. Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences* **27**, 258–281 (2023).
3. Pereira, F. *et al*. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications* **9**, 963 (2018).
4. Wang, Y., Ji, Q., Zhou, C. & Wang, Y. Brain mechanisms linking language processing and open motor skill training. *Frontiers in Human Neuroscience* **16** (2022).
5. Fuseda, K. *et al*. Impact of depressed state on attention and language processing during news broadcasts: EEG analysis and machine learning approach. *Scientific Report* **12**, 20492 (2022).
6. McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J. & Schütze, H. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences* **117**, 25966–25974 (2020).
7. Tang, J. *et al*. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* **26**, 858–866 (2023).
8. Sun, P., Anumanchipalli, G. K. & Chang, E. F. Brain2char: a deep architecture for decoding text from brain recordings. *Journal of neural engineering* **17**, 066015 (2020).
9. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* **5**, 1097–1107 (2023).
10. Jeong, J., Cho, J., Lee, B. & Lee, S. Real-time deep neurolinguistic learning enhances noninvasive neural language decoding for brain–machine interaction. *IEEE. Trans. Cybern*. (2022).
11. Artemova, E., Bakarov, A., Artemov, A., Burnaev, E. & Sharaev, M. Data-driven models and computational tools for neurolinguistics: a language technology perspective. Preprint at https://arxiv.org/abs/2003.10540 (2020).
12. Zock, M. AI at the crossroads of NLP and neurosciences. *Journal of Cognitive Science* **21**, 1–14 (2020).
13. Herff, C. *et al*. Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience* **9**, 217 (2015).
14. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
15. Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience* **23**, 575–582 (2020).
16. Sun, J., Wang, S., Zhang, J. & Zong, C. Towards sentence-level brain decoding with distributed representations. In *AAAI. Conf. Artif. Intell*., 7047–7054 (2019).
17. Sarker, I. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN comput. sci*. **2**, 420 (2021).
18. Teplan, M. *et al*. Fundamentals of EEG measurement. *Meas. Sci. Rev*. **2**, 1–11 (2002).
19. Dietrich, A. & Kanso, R. A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychol. Bull*. **136**, 822 (2010).

20. Grootswagers, T. *et al*. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Sci. Data.* **9**, 3 (2022).
21. Gifford, A. T., Dwivedi, K., Roig, G. & Cichy, R. M. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* **264**, 119754 (2022).
22. Telesford, Q. *et al*. An open-access dataset of naturalistic viewing using simultaneous EEG-fMRI. *Sci. Data.* **10**, 554 (2023).
23. Liu, J., Feng, K., Song, L. & Zeng, X. A visual EEG paradigm and dataset for recognizing the size transformation of images. In *2021 International Conference on Networking Systems of AI (INSAI)*, 171–175 (IEEE, 2021).
24. Hollenstein, N. *et al*. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Sci. Data.* **5**, 1–13 (2018).
25. Siok, W. T., Perfetti, C. A., Jin, Z. & Tan, L. H. Biological abnormality of impaired reading is constrained by culture. *Nature* **431**, 71–76 (2004).
26. Mou, X. *et al*. ChineseEEG: A Chinese linguistic corpora EEG dataset for semantic alignment and neural decoding. *Scidb* https://doi.org/10.57760/sciencedb.CHNNeuro.00007 (2024).
27. Mou, X. *et al*. ChineseEEG: A Chinese linguistic corpora EEG dataset for semantic alignment and neural decoding. *Openneuro* https://doi.org/10.18112/openneuro.ds004952.v1.2.0 (2024).
28. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics* (2019).
29. Xu, M., Li, D. & Li, P. Brain decoding in multiple languages: Can cross-language brain decoding work? *Brain and Language* **215**, 104922 (2021).
30. Hollenstein, N. *et al*. Decoding EEG brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience* **15** (2021).
31. Stehwien, S., Henke, L., Hale, J., Brennan, J. & Meyer, L. The Little Prince in 26 languages: Towards a multilingual neuro-cognitive corpus. In Chersoni, E., Devereux, B. & Huang, C.-R. (eds.) *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, 43–49 (European Language Resources Association, 2020).
32. Bhattasali, S., Brennan, J., Luh, W.-M., Franzluebbers, B. & Hale, J. The Alice datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 120–125 (European Language Resources Association, 2020).
33. Wang, S. *et al*. A synchronized multimodal neuroimaging dataset for studying brain language processing. *Sci. Data.* **9**, 590 (2022).
34. Li, J. *et al*. Le Petit Prince multilingual naturalistic fMRI corpus. *Sci. Data.* **9**, 530 (2022).
35. Wang, J., Chen, H.-C., Radach, R. & Inhoff, A. *Reading Chinese script: A cognitive analysis* (Psychology Press, 1999).
36. Peirce, J. *et al*. Psychopy2: Experiments in behavior made easy. *Behavior Research Methods* **51**, 195–203 (2019).
37. Chen, K. *et al*. A resource for assessing dynamic binary choices in the adult brain using EEG and mouse-tracking. *Sci. Data.* **9** (2022).
38. Gramfort, A. *et al*. MNE software for processing MEG and EEG data. *NeuroImage* **86**, 446–460 (2014).
39. Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M. & Robbins, K. A. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front. Neuroinform.* **9** (2015).
40. Li, A., Feitelberg, J., Saini, A. P., Höchenberger, R. & Scheltienne, M. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software* **7**, 4484 (2022).
41. Gorgolewski, K. *et al*. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* **3**, 160044 (2016).
42. Pernet, C. R. *et al*. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci. Data.* **6**, 1–5 (2019).
43. Appelhoff, S. *et al*. MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis. *Journal of Open Source Software* **4** (2019).
44. Fischl, B. R., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High—resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping* **8** (1999).
45. Grech, R. *et al*. Review on solving the inverse problem in EEG source analysis. *Journal of NeuroEngineering and Rehabilitation* **5**, 1–33 (2008).
46. Bi, Y. *et al*. The role of the left anterior temporal lobe in language processing revisited: Evidence from an individual with ATL resection. *Cortex.* **47**, 575–587 (2011).
47. Pobric, G., Ralph, M. A. L. & Jefferies, E. The role of the anterior temporal lobes in the comprehension of concrete and abstract words: rTMS evidence. *Cortex.* **45**, 1104–1110 (2009).
48. Fedorenko, E., Behr, M. K. & Kanwisher, N. G. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences* **108**, 16428–16433 (2011).
49. Po-Ching, Y. & Rimmington, D. *Chinese: A comprehensive grammar* (Routledge, 2015).
50. Wolf, T. *et al*. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Online, 2020).
51. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).
52. Appelhoff, S. *et al*. pybv – A lightweight I/O utility for the BrainVision data format.

Acknowledgements

This work was mainly supported by the MindD project of Tianqiao and Chrissy Chen Institute(TCCI), the Science and Technology Development Fund (FDCT) of Macau [0127/2020/A3, 0041/2022/A], the Natural Science Foundation of Guangdong Province(2021A1515012509), Shenzhen-Hong Kong-Macao Science and Technology Innovation Project (Category C) (SGDX2020110309280100), and the SRG of University of Macau (SRG2020-00027-ICI). We also thank all research assistants who provided general support in participant recruiting and data collection.

Author contributions

H.Wu, Q.Liu and X.Wang designed the study, H.Wu, Q.Liu and X.Wang, X.Mou, C.He, and L.Tan designed the experiments, [movie, Chinese text…], X.Mou, C.He and L.Tan, H.Liang and J.Zhang conducted the experiments, X.Mou, C.He, L.Tan, H.Liang, J.Zhang and J.Yu analyzed the results. X.Mou, C.He and L.Tan wrote the first draft. All authors checked the code, wrote the manuscript, reviewed the manuscript, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03398-7.

Correspondence and requests for materials should be addressed to Q.L. or H.W.

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

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