Received: 29 December 2025 | Revised: 15 April 2026 | Accepted: 15 April 2026

DOI: 10.1002/brx2.70046

RESEARCH ARTICLE

Toward general-purpose foundation models for electroencephalography: A unified data registry

Shengle Shi[1] | Yinglu Song[1] | Yong Wang[2] | Jiaqing Xiao[1] | Xinpeng Lin[1] | Heng Wang[1] | Zhibin Zhao[3] | Pengyu Wang[4] | Zitong Lu[5] | Xiaoli Li[1] | He Chen[1]

[1]School of Automation Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China

[2]Department of Rehabilitation Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China

[3]Department of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei, China

[4]Faculty of Medicine, School of Biomedical Engineering, Dalian University of Technology, Dalian, Liaoning, China

[5]Massachusetts Institute of Technology, McGovern Institute for Brain Research, Cambridge, Massachusetts, USA

Correspondence

He Chen, Xiaoli Li and Zitong Lu.
Email: chenhe@scut.edu.cn, xiaolili@scut.edu.cn and zitonglu@mit.edu

Funding information

This work was supported by the Strategic International Scientific and Technological Innovation Cooperation Initiative under the National Key R&D Program of China, Grant/Award Number: 2025YFE0205100; Funding by Science and Technology Projects in Guangzhou, Grant/Award Number: 2025A04J4000

Abstract

Electroencephalography (EEG) is widely used in cognitive neuroscience, clinical diagnosis, and brain–computer interfaces (BCIs). Recent work has begun to explore large-scale EEG pretraining for transferable representation learning, which requires diverse and well-organized EEG data across tasks and populations. Currently, public EEG datasets are scattered across platforms and publications and show substantial variation in experimental paradigms, recording settings, and metadata standards. This fragmentation makes large-scale discovery, integration, and reuse inefficient, particularly for foundation model pretraining. To address this, we systematically screened publicly available EEG datasets updated from 2020 to 2026 and constructed a unified EEG dataset registry tailored to scalable pretraining and benchmarking. The primary output of this work was a registry with structured metadata enabling efficient dataset discovery, filtering, and direct retrieval for EEG foundation model pretraining. We reviewed more than 900 publications and curated 827 eligible datasets, organized under a six-category taxonomy (cognitive, BCI, naturalistic, clinical, neuromodulation, and methodological). For each dataset, we recorded standardized metadata fields as reported, including task paradigm, device, channels, montage, sampling rate, number of participants, region, age, health status, license, data modalities, and label availability, together with embedded dataset and literature links to support direct retrieval. Using this curated inventory, we present a characterization of EEG resources, including domain imbalance and platform concentration, which highlights the difficulty of assembling corpora from sources. The registry offers centralized access and standardized descriptions, reducing the cost of discovery and cross-dataset alignment and supporting the pretraining and evaluation of EEG foundation models.

KEYWORDS

big data, data registry, electroencephalography (EEG), foundation models, open science

1 | INTRODUCTION

Electroencephalography (EEG) is a noninvasive technique that records brain electrical activity via electrodes placed on the scalp.[1] Owing to its high temporal resolution, relatively low cost, and ease of deployment, EEG has been widely used in cognitive neuroscience research,[2–6] clinical diagnosis and prognostic evaluation,[7–11] and state monitoring and feedback signal acquisition in closed-loop neuromodulation systems.[12–14] In brain–computer interfaces (BCIs), EEG is commonly employed for external device control and cognitive state decoding.[15–20] It has also been applied for the recognition of states such as attention,[21] fatigue, and workload.[22,23] Recently, the rapid development of portable and wearable EEG devices[24] has enabled researchers to acquire EEG signals in real-world scenarios such as physical movement and driving,[25,26] thereby further expanding the application scope of EEG.[27–30]

As application contexts broaden, the quantity and variety of public EEG datasets have steadily increased. These datasets are disseminated across several platforms, including OpenNeuro, Zenodo, Figshare, OSF, ScienceDB, and PhysioNet. They encompass a wide range of task paradigms and subject groups, and some further include multimodal synchronous recordings and detailed annotations. Nevertheless, current public resources are highly fragmented, as datasets differ markedly in experimental designs, recording equipment, data formats, and annotation definitions, thereby raising the costs associated with data search, comparison, and reuse. In deep learning driving advances in large-scale representation learning, how to systematically identify and integrate available EEG data resources has become a prerequisite for related research, including foundation model pretraining.

At present, deep learning–based artificial intelligence approaches are extensively used in computer vision,[31,32] natural language processing, medical imaging, and related domains.[33–36] This progress has fostered the emergence of foundation models pretrained on large-scale and heterogeneous datasets.[33,37] For instance, GPT-3 exhibits robust generalization under few-shot and zero-shot conditions through autoregressive pretraining on large text corpora, whereas CLIP employs contrastive pretraining on large collections of image–text pairs to learn cross-modal alignment, exemplifying a paradigm that builds on universal representation pretraining followed by downstream task adaptation. Central to this paradigm is the ability to learn general and transferable representations that perform well across various downstream tasks. For example, BERT's pretraining strategy enables effective adaptation to tasks such as question answering, text classification, and sequence labeling.[34] Similarly, visual self-supervised methods such as masked autoencoders (MAE) provide strong initial representations that benefit image classification, object detection, and segmentation tasks.[38] In neuroscience, recent work has focused on learning shared representations across different types of brain signals. One example is BrainOmni, which combines EEG and magnetoencephalography (MEG) data and demonstrates the potential for learning that generalizes across tasks, devices, subjects, and signal modalities.[39] In particular, BrainOmni highlights effective alignment between EEG and MEG representations. In parallel, large-scale EEG-based models such as LaBraM show that learned representations can transfer reliably across different subjects and experimental tasks.[40]

However, such explorations have not yet formed a widely accepted paradigm in EEG research. Overall, most models in the EEG domain are still designed and trained around a single task or a single dataset; for example, SleepEEGNet for sleep staging represents a classic task-specific modeling approach.[41] Recent foundation model studies have also explicitly shown that existing methods often rely on isolated, modality- or dataset-specific model paradigms, thereby limiting cross-domain scalability and reuse. Simultaneously, LaBraM further emphasizes that, compared with the existing approaches designed for specific BCI tasks, it shows a clear advantage in terms of generalization capability.[40] Existing pretraining frameworks for general-purpose EEG foundation models remain limited, as they lack universal training strategies capable of robust generalization across tasks, devices, and subjects. Recent research efforts have started to tackle these issues. For example, EEGPT aims to learn general and reliable representations to mitigate cross-subject variability and channel mismatch,[42] whereas BIOT focuses on cross-dataset learning to address channel inconsistency, variable sequence lengths, and missing data.[43] LaBraM similarly presents itself as an early

Key Points

What is already known?

- Public electroencephalography (EEG) datasets have become increasingly important for large-scale pretraining and general-purpose foundation model research.
- However, these datasets are scattered across repositories and publications, with substantial heterogeneity in paradigms, acquisition settings, and metadata formats.
- This fragmentation makes dataset discovery, comparison, and cross-dataset reuse difficult.

What does this study add?

- It introduces a unified public EEG warehouse that consolidates 827 datasets and more than 130,000 participants in a single structured resource.
- It provides rich, standardized metadata to support efficient dataset discovery, filtering, comparison, and retrieval.
- It offers a practical infrastructure for scalable EEG pretraining, benchmarking, and foundation model development.

exploration of large-scale EEG pretraining, highlighting a substantial gap in both scale and capability when compared with foundation models in vision and language.[40]

The identified gap may primarily stem from limitations at the data level. Currently, there is no unified, centrally aggregated, and standardized EEG data resource. Consequently, these data are challenging to leverage at scale and remain inadequate for the data acquisition, alignment, and organization necessary for large-scale pretraining. TUH-EEG points out that although current clinical EEG datasets are large in scale, the portion that can be directly used for machine learning remains limited and highly heterogeneous.[44] BrainOmni also indicates that cross-device variability and the lack of standardization directly affect a model's ability to generalize across devices.[39] EEG-BIDS incorporates EEG data into the BIDS organizational framework, highlighting the urgent need for unified data standards to improve data sharing and utilization efficiency.[45]

At present, researchers mainly obtain EEG datasets through two primary approaches. The first approach involves independently searching for and downloading datasets from public platforms and the literature based on specific research questions. For example, researchers can filter datasets by task, population, or acquisition conditions on platforms such as PhysioNet and OpenNeuro, or trace data sources through data availability statements in publications.[46,47] The second approach relies on dataset reviews focused on particular subdomains, such as surveys of publicly available motor imagery (MI) datasets in BCI research and overviews of clinical epilepsy EEG datasets.[48,49] Such reviews are typically of high quality, but they tend to focus on a single domain and are unable to simultaneously cover data needs across multiple tasks and paradigms. Therefore, from the perspective of foundation model pretraining, existing approaches to EEG data acquisition remain relatively fragmented. Although large-scale clinical datasets (e.g., TUH-EEG) and widely used task-specific datasets (e.g., DEAP) are available, they exhibit substantial differences in task definitions, acquisition protocols, channel configurations, and annotation schemes.[44,50] Conversely, the lack of a unified data entry point and standardized organizational schemes increases the cost of data alignment and organization.[45]

Given the abovementioned context, constructing an EEG data repository tailored to the data requirements of foundation models is crucial. Unified aggregation and standardized descriptions help improve the discoverability and reusability of public datasets, while reducing the cost of data cleaning and organization in large-scale pretraining. This objective is aligned with the FAIR principles and data standardization practices promoted by EEG-BIDS and OpenNeuro.[45,46,51] Motivated by this consideration, we systematically screened and curated publicly available EEG resources updated after 2020 and established a unified classification framework. To enhance transparency and reproducibility, we adopted a question-oriented taxonomy that organizes datasets by their primary scientific intent and data-generation context, rather than by hosting platform or surface task names. We selected this principle because it captures stable sources of domain shift relevant to EEG foundation-model pretraining, including differences in task objective and output type (e.g., decoding/control vs. mechanism probing), participant population (healthy vs. clinical), acquisition context (highly controlled paradigms vs. ecological settings), and the presence of external interventions that can alter signal distributions. Accordingly, we define six categories: BCI (datasets released for decoding/control), cognitive (controlled paradigms probing core cognitive processes), clinical (datasets including formally diagnosed patient cohorts or clinically defined disorders), naturalistic (ecological or weakly constrained real-world scenarios), neuromodulation (EEG collected under explicit stimulation/training interventions), and methodological (datasets released for artifact/quality studies, or benchmarking). Using this taxonomy, we screened and compiled over 800 publicly available EEG datasets and further organized each category into refined subcategories. An overview of this taxonomy and the corresponding distribution of datasets is shown in Figure 1.

For each dataset, we systematically recorded a multi-field set of metadata as reported by the original sources, including task paradigm, device, channels, montage, sampling rate, number of participants, region, age, health status, license, data modalities, label availability, hosting platform, publication year, access links, and associated references.

When specific metadata items were not explicitly reported in the original dataset repository or publication, the corresponding fields were recorded as N/A. This work aimed to provide a broad-coverage and structured data access point rather than conducting in-depth methodological evaluations of individual datasets. By explicitly organizing these metadata fields and access pathways, the registry is anticipated to support efficient dataset discovery, filtering, compatibility assessment, and scalable dataset aggregation for subsequent EEG foundation model pretraining. It is publicly available at Zenodo (https://zenodo.org/records/18815016).

2 | METHODS

This study sought to present a systematic curation of publicly available EEG datasets. The overall workflow was conducted in accordance with the PRISMA-ScR guidelines to ensure process transparency and traceability.[52] As the study focused on datasets rather than traditional publications, we defined a three-stage workflow: (1) search across multiple data repositories; (2) manual screening based on dataset descriptions; (3) post-download assessment of data usability and metadata completion.

2.1 | Search strategy

Publicly available EEG datasets were used as the retrieval targets, following a search strategy centered on dataset platforms. The search platforms included both domain-specific

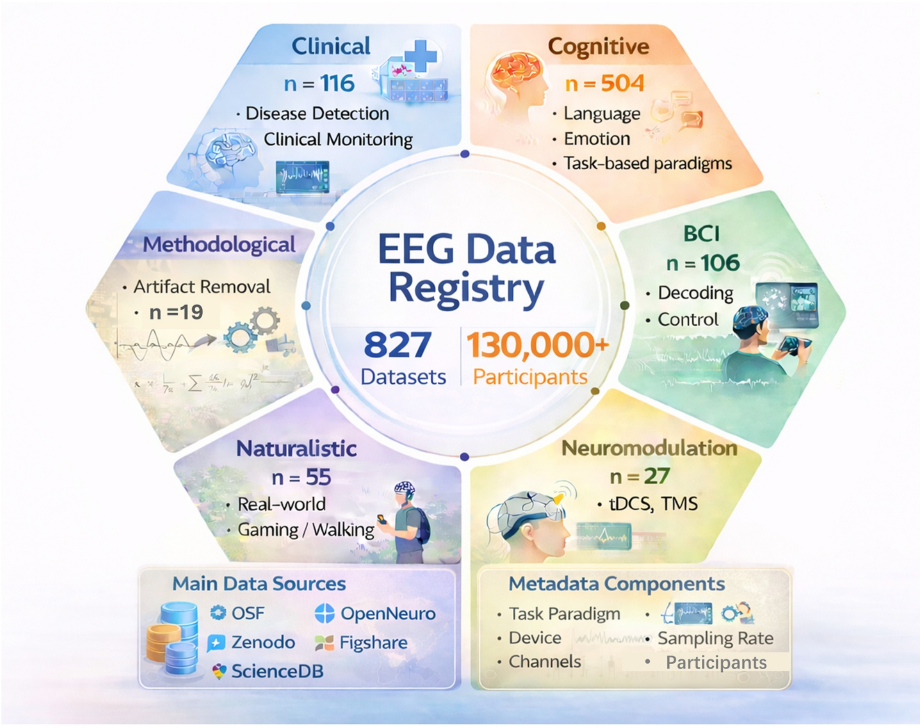

FIGURE 1 Overview of the EEG data registry. EEG, electroencephalography.

repositories for neural and physiological signals and general scientific data hosting platforms, such as OpenNeuro, PhysioNet, OSF, Zenodo, Figshare, ScienceDB, Mendeley Data, and IEEE DataPort, with specific URLs provided in Table 1.

On each platform, the terms "EEG" and "electroencephalography" were employed as primary keywords to search dataset titles, keywords/tags, and searchable description fields (e.g., descriptions, abstracts, and README summaries) without imposing any language constraints. The last update time displayed on the data portal websites was used as the temporal reference, and dataset records with update dates from 2020 to 2026 were included. When an explicit update timestamp was not provided by a platform, alternative rules were applied sequentially to determine the timestamp, with the source documented. First, the most recent date from the platform's version records or release history was used. Second, the version or publication date provided in the dataset citation information was adopted. Finally, a verifiable publication date from associated research articles or data descriptor papers was used. The search process was conducted from Sep 6, 2025 to January 13, 2026. For continuously updated institutional databases, we additionally recorded their update nature and subcorpus structure to improve transparency and reduce omissions within the defined search window.

The following search string was used for data portals:

(TITLE: ("EEG" OR "electroencephalography") OR KEYWORDS/TAGS: ("EEG" OR "electroencephalography") OR DESCRIPTION: ("EEG" OR "electroencephalography"))

TABLE 1 List of the EEG repositories screened for suitable datasets.

Repository	URL
Neurology datasets repositories	
Openneuro	https://openneuro.org/
PhysioNet	https://physionet.org/
G-node	https://gin.g-node.org/
Databases of data-sharing portals	
OSF	https://osf.io/
Zenodo	https://zenodo.org/
Figshare	https://figshare.com/
Mendeley data	https://data.mendeley.com/
Dryad	https://datadryad.org/
ScienceDB	https://www.scidb.cn/en

2.2 | Inclusion and exclusion criteria

Dataset records served as the inclusion units, as opposed to scholarly publications that only report on datasets. Datasets were eligible for inclusion if they: (i) contained EEG signal data suitable for analysis; (ii) provided minimally sufficient, traceable citation information; and (iii) offered access to essential metadata required for reuse via the dataset page and/or related publications. Because metadata detail varies substantially across platforms, entries with incomplete information on the record page were permitted to have key metadata completed during the secondary screening by consulting linked articles or author information.

The inclusion criteria for the datasets included the following:

- The dataset record clearly referred to EEG or electroencephalography in its title, keywords, or description;
- The dataset record page was accessible, and the resource was an independent dataset rather than solely code, tutorials, tools, or sample files;
- The dataset's last updated date on the data portal was within 2020–2026, according to the specified time-window criteria;
- The dataset included obtainable or verifiable EEG signal data, either in a unimodal form or within multimodal synchronized recordings, as long as EEG data were present;
- The record included traceable citation information, such as a DOI, persistent identifier, or standardized platform citation, enabling data traceability and proper referencing;
- The dataset offered at least minimal information to support reuse—such as task background, subject details, file organization, or annotations—or allowed essential metadata to be supplemented via related publications.

Conversely, the exclusion criteria for datasets were:

- The dataset record did not mention "EEG" in its title, keywords, or description, and no evidence suggested that EEG data were included;
- The resource was not a standalone dataset—being limited to code, software, tutorials, or sample/test files—and did not contain analyzable recorded data;
- The dataset page was inaccessible, empty, or not publicly available;
- The dataset was released only through journal supplementary materials in the referenced articles without an independently hosted dataset record (e.g., repository/institutional portal page) or a stable, reproducible access pathway;
- The last update time shown on the data portal was outside the 2020–2026 time window;
- After the description or downloaded files were checked, the dataset was confirmed to exclude EEG, such as containing only MEG, fMRI, behavioral/questionnaire, or eye-tracking data;
- Essential metadata were absent and could not be recovered from related publications, making the task design, participant characteristics, or recording conditions unclear;
- The downloaded data files were corrupted, could not be read, or lacked required accompanying files (e.g., event annotations, channel information, or participant metadata), hindering reuse.

After the aggregation of aggregating dataset records retrieved from all platforms, deduplication was performed. The deduplication targets mainly included the following:

- Duplicate releases across platforms;
- Different versions or mirrored entries of the same dataset within a single platform;
- Records with highly similar titles and descriptions pointing to the same data source.

Deduplication decisions were primarily based on DOIs, persistent links, project homepages, author information, and version metadata. Records suspected to originate from the same source but lacking sufficient information were retained and further verified during the secondary screening stage. In the rescreening stage, dataset usability was assessed at the download and file levels, with emphasis on the following:

- File completeness and decompressibility;
- Readability and parsability of EEG data files;
- Transparency of data organization and formats (e.g., BIDS and, EDF/BDF/SET/MAT);
- Provision of essential metadata for reuse, including channel layouts, event markers, participant information, and README or data dictionary files.

For datasets with insufficient information on the record page, associated journal articles, preprints, or data descriptor papers were retrieved based on the provided citation details, author information, or project clues to supplement key metadata such as task paradigms, participant characteristics, and acquisition protocols. Ultimately, only datasets that actually contained usable EEG data, provided traceable citation information, and met the minimum reuse requirements of this study were included. The numbers of records identified, screened, and included or excluded were summarized using a PRISMA flow diagram (Figure 2).

2.3 | Extracted information

2.3.1 | Metadata extraction and verification

For each dataset ultimately included, we used a predefined structured extraction form to collect metadata from the dataset repository pages and associated publications. When information on the dataset repository page was insufficient, descriptions of acquisition protocols, experimental tasks, and data organization provided in the associated papers were used as supplementary sources. In cases of discrepancies between the two sources, verifiable data file structures and contents were prioritized as the primary reference, and any differences along with their evidence sources were documented. For each dataset ultimately included, metadata were extracted with emphasis on the following:

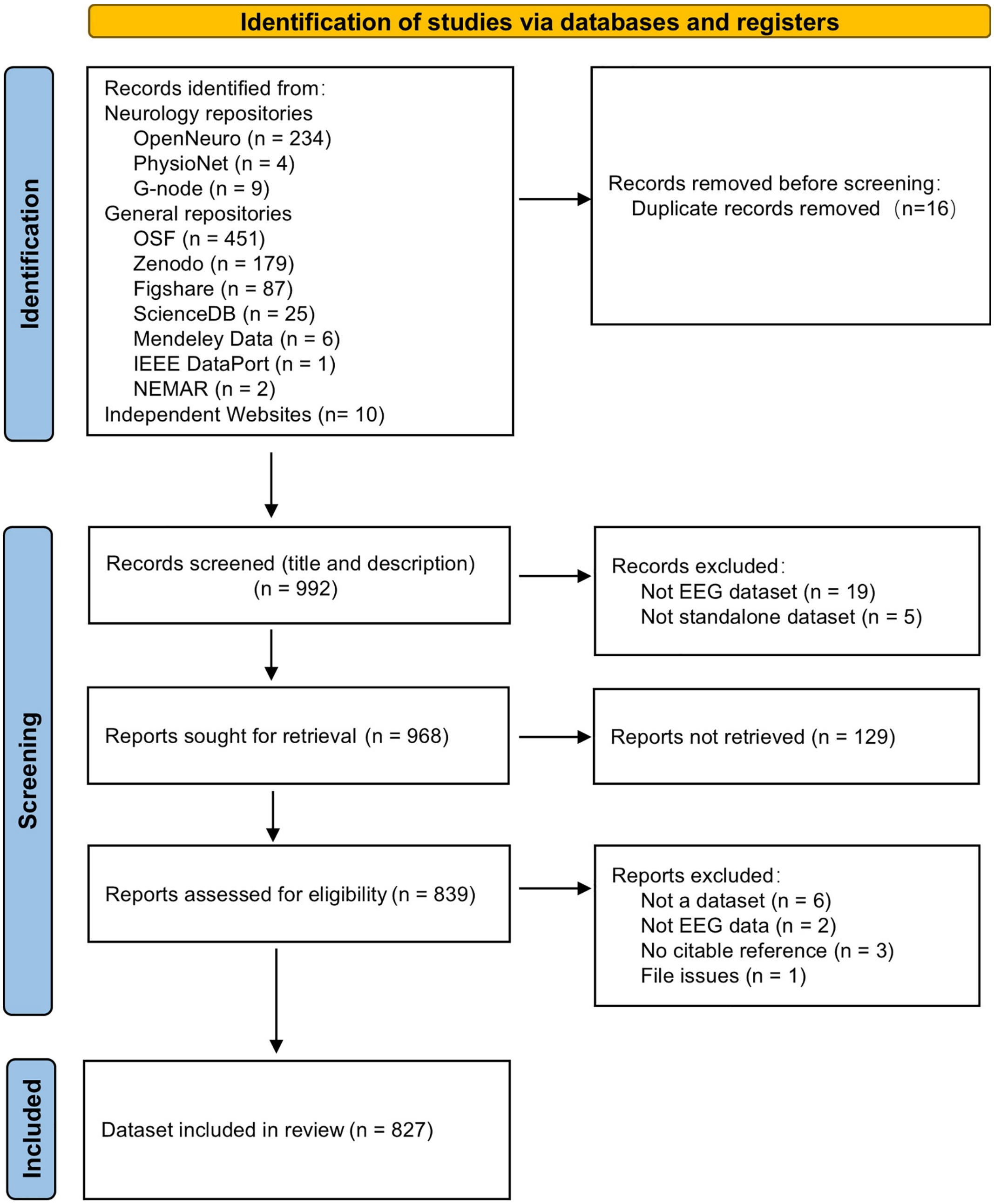

FIGURE 2 Dataset selection for inclusion using the PRISMA framework.

- Citation and traceability information: DOI (if available), standardized citation entry or persistent link, unique dataset URL, and references to associated publications;
- Access and release information: access conditions (e.g., open, restricted, application-based) hosting platform and publishing organization, last update date recorded by the portal, and version information when available;
- Task and topic annotations: primary categories and subcategories;
- Subject and sample characteristics: population type, sample size and grouping (when reported), and number of classes or experimental conditions, if available;
- Data modalities and recorded content: modality composition, and EEG type (e.g., resting-state or task-based);
- Experimental and annotation details: number of sessions, recording duration or number of trials, annotation sources and formats, and alignment of event information with EEG signals;
- Acquisition devices and parameters: device type, number of channels, electrode layout, and sampling rate;
- Documentation and reuse support: presence of structured documentation artifacts, including paradigm descriptions, data dictionaries (variable definitions and directory structure), and example code or analysis scripts.

2.3.2 | Taxonomy assignment

For each curated dataset, we determined its category by consulting the dataset landing page on the hosting platform (repository description/README/portal metadata) and, when needed, the associated publications. Category assignment was anchored to the dataset's explicitly stated primary objective and data-generation context, rather than to the hosting platform or surface task names alone. In addition to the six top-level categories, we further recorded refined topic tags to reflect major subdomains. Each dataset in the registry was assigned a unique numerical identifier (ID) corresponding to its entry in the master table; the ID served as a stable internal reference and did not encode category, platform, or temporal information.

Because many public EEG datasets naturally exhibit overlapping characteristics, we adopted a consistent rule to ensure corpus-level interpretability: each dataset received a single primary category for global statistics, while other attributes (e.g., paradigm keywords, auxiliary modalities, and intervention presence) were preserved as structured metadata fields. Datasets including formally diagnosed patient cohorts or clinically defined disorders were categorized as clinical, even when canonical cognitive paradigms were used during acquisition. For emotion-related datasets, when the dataset was explicitly released for emotion decoding/recognition performance (e.g., supervised classification), we categorized it as BCI; when the primary goal was to probe mechanistic emotion processing under controlled experimental conditions, we categorized it as cognitive. Finally, the presence of continuous or naturalistic-looking stimuli did not by itself determine naturalistic categorization; we assigned a dataset to naturalistic only when its primary data-generation context was ecologically oriented and long-form, where trial-based experimental control was not the organizing principle of the recording. Illustrative borderline cases for overlap resolution are provided in the Discussion section below.

3 | RESULTS

3.1 | Overview

Following initial screening, 968 dataset records were sought for retrieval (Figure 2). After data access and eligibility assessment, 827 datasets satisfied all inclusion criteria and were included in the final analysis. The included datasets were categorized according to a predefined classification framework (Figure 3A). Overall, the cognitive datasets constituted the largest category (n = 504), followed by the clinical (n = 116) and BCI datasets (n = 106). Additional categories included the naturalistic (n = 55), neuromodulation (n = 27), and methodological datasets (n = 19).

In terms of data hosting platforms (Figure 3B), the included datasets were predominantly concentrated within a few major open science and neuroscience repositories, with OSF contributing the largest number (n = 353), followed by OpenNeuro (n = 173), Zenodo (n = 147), Figshare (n = 78), and ScienceDB (n = 25). Collectively, these four platforms accounted for 776 out of 827 included entries (93.8%), serving as the principal carriers of public EEG data resources in this study. The remaining datasets were scattered across various long-tail platforms and standalone websites, such as G-Node (n = 9), independent sites (n = 19), Dryad (n = 8), Mendeley Data (n = 6), and PhysioNet (n = 3). A limited number of datasets were also hosted on platforms including NEMAR, GigaDB, GitHub, IEEE DataPort, Kaggle, OpenAIRE, and Synapse ($n \leq 2$). Notably, the platform statistics presented in this section pertain solely to the source distribution of the included sample (n = 827) and do not indicate the overall volume of EEG datasets available on each platform. Therefore, the hosting platforms and corresponding links were explicitly recorded for each dataset entry to support future expanded searches and data reuse. The full multi-field registry table used for these statistics is available in filterable Excel formats at Zenodo (https://zenodo.org/records/18815016).

Furthermore, we analyzed the correspondence between major hosting platforms and major dataset categories. Considering the limited robustness of category distributions on small-sample platforms, we restricted the comparison to platforms with more than five dataset entries (OSF, OpenNeuro, Zenodo, Figshare, G-Node, Dryad, and Mendeley) and excluded records hosted on independent websites. As shown in Figure 4, pronounced systematic differences were observed in dataset category composition across the platforms. OSF was substantially dominated by the cognitive datasets (n = 275), with the remaining categories (clinical, BCI, naturalistic, neuromodulation, and methodological) contributing relatively small proportions, resulting in a strongly single-dominant structure. Similarly, OpenNeuro was dominated by the cognitive datasets (n = 105), but its comparatively larger shares of clinical (n = 33) and BCI datasets (n = 18) indicate a greater emphasis on clinically and technically oriented research. In contrast, Zenodo and Figshare contributed more prominently to the BCI datasets (n = 39 and n = 29, respectively); Zenodo also hosted a substantial number of the clinical datasets (n = 28), whereas Figshare included relatively fewer clinical datasets (n = 10), reflecting differences in research orientation. The platforms with smaller numbers of records, such as G-Node, Dryad, and Mendeley, were almost entirely concentrated in the cognitive datasets (n = 8, 6, and 3, respectively) with only few clinical or BCI dataset entries and limited category coverage.

3.2 | Datasets under various conditions

3.2.1 | Cognitive datasets

Among the datasets included in this study, the cognitive datasets formed the largest group (n = 504), covering diverse task paradigms ranging from sensory processing to cognitive

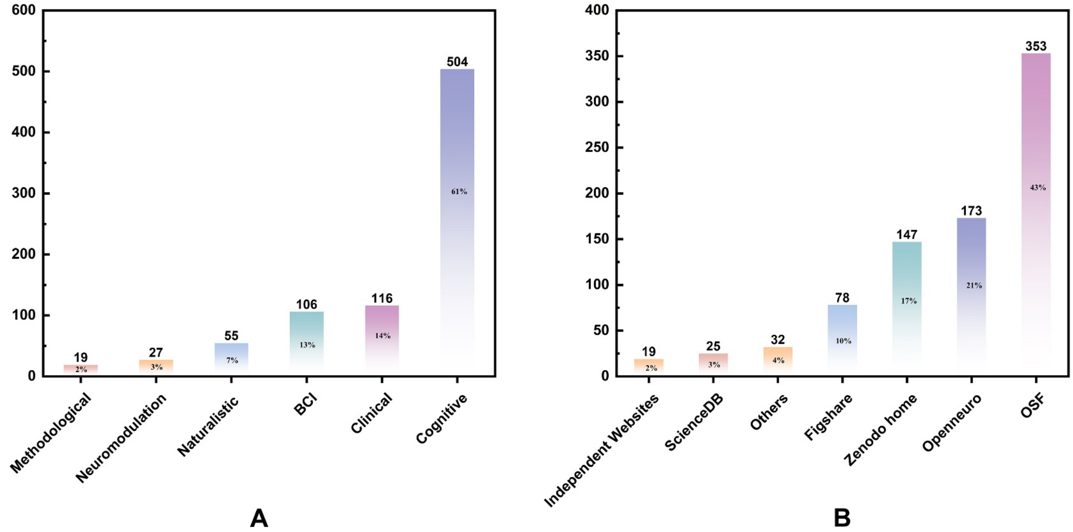

FIGURE 3 Structural distribution of the EEG datasets across research categories and data platforms. (A) Number of datasets included in this study for each research category. (B) Number of datasets included in this study from each data platform.

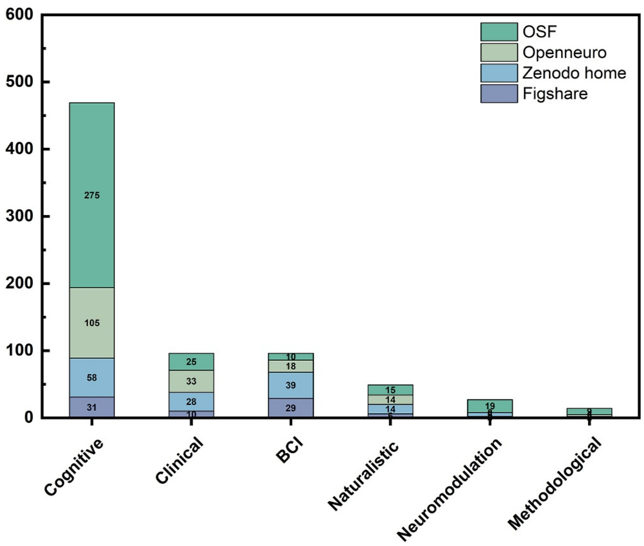

FIGURE 4 Distribution of the major EEG datasets across the major repository platforms.

control. The datasets were mainly concentrated in four subcategories—perception (n = 100), learning and memory (n = 90), language processing (n = 66), and attention (n = 60)—collectively comprising about two-thirds of the cognitive datasets. This pattern indicates that public EEG cognitive datasets are still largely based on event-related classical paradigms, usually offering well-defined event annotations and thus facilitating reuse.[53] Furthermore, the datasets on executive function (n = 35) and emotion processing (n = 40) offered more complex tasks involving cognitive control and emotional regulation.[54] The social cognition datasets (n = 30) were mainly centered on interpersonal interaction tasks and cooperative or competitive scenarios.[55] The developmental and aging datasets (n = 19) spanned infant, child, adolescent, and older adult populations to characterize cognitive differences across developmental stages.[56] Beyond these major subcategories, several minor domains were also represented, including consciousness (n = 12), decision making (n = 19), sleep (n = 13), metacognition (n = 9), and spatial cognition (n = 2). Because the cognitive datasets were numerous and mainly concentrated in a few subcategories, the main text emphasizes the six most representative subdomains and synthesizes their typical task paradigms and labeling strategies. The full records of all remaining subcategories with their access links are available at Zenodo (https://zenodo.org/records/18815016). Across these cognitive datasets, experimental data were generally organized by trials, stimulus types, and time-locked events, featuring clear event markers and well-defined epoch structures, which make them particularly suitable for event-related potential (ERP) and time–frequency analyses and facilitate data reuse and cross-study comparison.

Perception

The perception category (n = 100) was largely dominated by the visual (n = 61) and auditory processing datasets (n = 16), jointly comprising approximately 76% of this category. The auditory processing datasets primarily used natural sounds, speech, and music as stimuli, covering speech materials, musical segments, and environmental sounds, and included acoustic characteristics such as frequency, pitch, intensity, and temporal structure.[57,58] Common experimental paradigms include the oddball paradigm, auditory steady-state response, and sound source localization and separation tasks under multi-source conditions.[57,58] These paradigms are concretely represented in the public datasets summarized herein. For instance, dataset #438 examined auditory processing differences across speech, music, and environmental sound stimuli[59]; dataset #535 provided a systematic comparison of expectancy violation processing under various oddball and local–global paradigms[60]; and dataset #859 examined sound source localization and spatial auditory processing through controlled manipulation of sound source locations.[61]

The visual processing datasets encompassed a broader range of stimuli and paradigms, including static images, naturalistic scenes, face or object stimuli, and perception tasks defined by parametric visual features such as orientation, spatial frequency, and contrast.[62–65] Typical paradigms comprised visual oddball, rapid serial visual presentation (RSVP), one-back, predictive cueing, and visual search and classification tasks. For instance, dataset #103 used an RSVP paradigm to present multi-class object images and recorded infants' EEG responses under rapid visual stimulation to study early visual processing[66]; dataset #246 investigated the influence of predictive cues and face stimuli on ERPs to probe visual expectation and repetition effects[63]; and dataset #420 applied an oddball paradigm that varies shape and examined distributions to elicit vMMN and examine visual processing under limited-attention conditions.[67]

In addition, the perception category comprised a limited number of datasets addressing multisensory integration (e.g., audio–visual or olfactory–visual, such as datasets #12 and #62),[68,69] time perception (e.g., dataset #85),[70] somatosensory perception,[71] and numerical cognition; key details for these datasets are available at Zenodo (https://zenodo.org/records/18815016).

Memory and learning

The memory and learning category was largely dominated by the working memory ($n = 41$) and long-term memory datasets ($n = 22$), jointly comprising 70% of this group. The working memory datasets focused on information encoding, maintenance, and updating, mainly employing visual stimuli, with certain tasks integrating auditory or cross-modal inputs. Typical paradigms comprised n-back, Sternberg, and DMS tasks, which systematically manipulate memory load, interference, and delay intervals to assess working memory capacity and maintenance processes.[72–74] For instance, dataset #230 examined the rapid clearance of memory representations after goal completion in a visual working memory paradigm, highlighting the goal-oriented nature of working memory under changing task demands.[75] Conversely, dataset #688 associated visual working memory items with specific actions and captured parallel neural dynamics of visual and motor selection during memory-guided behavior, probing how mnemonic representations are converted into actions.[76]

The long-term memory datasets were primarily composed of episodic memory datasets, emphasizing encoding and retrieval processes across temporal delays and the binding of memory contents to semantic or contextual cues. Typical paradigms comprised item recognition, source memory, and free recall. For instance, dataset #213 used recognition paradigms together with multivariate classification to study the continuity of neural states between encoding and retrieval,[77] while dataset #745 (PEERS) included various free recall and recognition paradigms with data from several hundred participants, facilitating analyses of neural representations and individual variability in long-term memory processes.[78]

Additionally, a limited number of datasets focused on learning and neural plasticity are available, with specific entries listed in Zenodo (https://zenodo.org/records/18815016).

Language processing

The language processing category ($n = 66$) centered on neural representations of linguistic information across perception, comprehension, and control, with stimuli primarily delivered through visual and auditory language inputs, including words, sentences, and continuous speech. Typical tasks comprised lexical and semantic processing, syntactic operations (e.g., violation detection), and naturalistic reading and auditory comprehension paradigms, capturing the temporal dynamics and hierarchical organization of language processing.[79,80] For example, dataset #323 recorded bilingual EEG data during cognitive control tasks while accounting for differences in language experience, revealing how language use patterns modulate spectral and ERP indices related to language control[5]; In contrast, dataset #639 recorded multichannel EEG responses of infants to natural speech syllables and applied multivariate analyses to reveal the decompositional and combinatorial encoding of consonant and vowel features, providing a data foundation for studying early language perception.[81]

Attention

The attention datasets primarily investigated how information is selected, enhanced, or suppressed under conditions of multiple stimulus competition or uncertainty, with stimuli mainly presented in visual and auditory modalities. Typical paradigms included comprise spatially cued cue–target tasks,[82]target–distractor search paradigms,[83]and temporal attention designs in continuous stimulus streams,[84] which correspond to the allocation and modulation of attention across spatial, feature, and temporal dimensions, respectively. For instance, dataset #441 varied the number of task-relevant features to examine global shifts in attentional strategies under uncertain conditions and corresponding cortico-thalamic activity changes.[85] Conversely, dataset #447 integrated subjective reports with high-density EEG to delineate local slow-wave activity associated with attentional lapses caused by internal attentional shifts or interruptions, offering a valuable resource for investigating sustained and internal attention states.[86]

Executive control

The executive control datasets mainly investigated conflict, error monitoring, and behavioral adjustment in goal-directed actions, centering on inhibition, updating, and task-switching mechanisms. Experimental designs commonly employed classic conflict and inhibition paradigms, such as Stroop, Flanker, Simon, Go/No-go, and stop-signal tasks, as well as rule-switching designs to manipulate control demands and conflict levels.[87–89] For instance, dataset #238 used an audiovisual ventriloquism illusion paradigm to induce cross-modal spatial conflict and measured reaction time and fronto-central theta activity to examine conflict processing during multisensory integration[90]; dataset #383 combined rule selection and stop-signal tasks and used EEG representational similarity analysis to track changes in action-related representations, thereby investigating modulatory processes in action inhibition.[54]

Emotion processing
The motion processing datasets primarily focused on the perception, evaluation, and regulation of emotions, using mainly visual and auditory stimuli, with some tasks manipulating emotional valence and arousal via pain or fear conditioning. Typical paradigms comprised emotional face and image recognition, emotion induction and regulation tasks, and emotion–cognition interaction paradigms, capturing emotional modulation of perception, attention, and decision processes.[91,92] For example, dataset #229 used emotional face stimuli to compare early emotion-related ERP components across different populations, facilitating the study of individual differences in emotional reactivity,[93] whereas dataset #449 integrated emotional face stimuli with feature-based attention paradigms to assess how attentional allocation modulates emotion-related ERP components.[94]

Within the cognitive datasets, in addition to the high-frequency subcategories discussed perception, learning and memory, language processing, attention, executive function, and emotion processing, our database also included several smaller yet thematically well-defined cognitive subcategories, such as social cognition, development and aging, consciousness, decision making, sleep, and metacognition. Due to space limitations, these subcategories are not discussed individually herein; the corresponding datasets and key information are systematically compiled and available at Zenodo (https://zenodo.org/records/18815016).

3.2.2 | BCI datasets

Of the included datasets, the BCI datasets ($n = 106$) formed a distinct major category, whose aims ranged from fundamental neural signal decoding to real-world human–machine interaction and assistive control. Overall, the BCI datasets were divided into three subcategories: The BCI control datasets ($n = 82$), BCI cognitive decoding datasets ($n = 20$), and BCI passive monitoring datasets ($n = 4$). BCI control datasets markedly dominated this category, consistent with the long-standing focus of BCI research on MI, ERP (e.g., P300), and steady-state visual evoked potential (SSVEP) as core control paradigms, indicating that current public EEG BCI datasets primarily focused serve command-level control and real-time interaction needs.[95,96] In contrast, the BCI cognitive decoding datasets primarily focused on the direct decoding of internal cognitive states such as emotion, attention, or language intent[97]; BCI passive monitoring datasets focused on the assessment and tracking of continuous states such as stress, workload, and fatigue.

Notably, the tasks involving emotion, language, and attention in the BCI datasets, despite superficial similarities in stimuli or task structures to the corresponding cognitive subcategories, were fundamentally different in their research goals and design rationale. In the cognitive datasets, these tasks primarily served to explain cognitive mechanisms and test theoretical hypotheses, focusing on neural processing, condition contrasts, and interpretability. In the BCI datasets, the corresponding cognitive processes were explicitly defined as brain states to be decoded, and experimental designs placed greater emphasis on engineering metrics such as signal stability, cross-subject generalization, and decoding accuracy. Consequently, the BCI datasets exhibited highly engineered and standardized experimental structures: stimulus–command mappings were strictly fixed; label definitions were centered on decoding consistency rather than psychological process differentiation, and system feedback mechanisms were fully embedded within the experimental workflow. This high degree of standardization rendered such datasets more suitable for algorithm performance benchmarking, cross-subject generalization evaluation, and online BCI system validation, rather than for direct modeling of cognitive mechanisms or theoretical hypothesis testing.

BCI control
The BCI control datasets were designed to convert EEG signals into control commands with an emphasis on real-time operation and accuracy, constituting the most mature and largest-scale area of BCI research. Such datasets primarily focused on several canonical control paradigms, including MI, SSVEP, and ERP-based control based on P300 or RSVP paradigms.[98] The MI datasets utilized μ/β rhythms induced by imagined limb movements to produce continuous or discrete commands, and were widely used in cursor control, rehabilitation, and robotic manipulation.[99,100] For instance, dataset #596 used standing and sitting lower-limb MI tasks to supply MI control data for rehabilitation and assisted gait contexts[101]; Conversely, dataset #613 presented noninvasive EEG-based multi-DOF robotic arm control, illustrating the potential of MI control in real-world interaction and prosthetic manipulation.[100] The SSVEP datasets elicited steady-state responses through periodic visual stimulation to support multi-target selection and high information transfer rates, and were widely used in spellers, menu selection, and wearable interactive systems[102]; In contrast, dataset #541 extended SSVEP control into head-mounted AR contexts and compared binocular congruent and incongruent frequency coding approaches, offering practical routes for wearable and portable SSVEP applications.[103] The P300 datasets commonly used oddball or speller paradigms to implement attention-driven command input by detecting target-related ERPs.

BCI cognitive decoding
The BCI cognitive decoding datasets were mainly used to identify and characterize cognitive states from EEG data using tasks that involve attention, language, and emotion processing. Compared with command-based paradigms such as MI and SSVEP, these datasets relied less on periodic

stimulation or fixed commands and more on natural or semi-natural contexts—such as multi-source auditory scenes, natural speech materials, or emotion-eliciting stimuli—to obtain neural features for discriminating cognitive states. The existing public datasets mainly concentrated on three tasks: attention decoding, language decoding, and emotion decoding. In attention decoding, the datasets often targeted attentional focus in multi-target or multi-source environments. For example, dataset #607 employed spatial auditory attention tasks, recording EEG data under complex audiovisual conditions to distinguish participants' attentional directions and analyze factors affecting decoding stability.[104] In language decoding, dataset #564 provided large-scale inner-speech EEG data spanning multiple semantic categories for studying language and semantic discrimination.[105] The emotion-decoding datasets typically elicited emotional states using images, videos, or imagery tasks. For example, dataset #532 integrated multiple emotion-elicitation contexts to support discrimination and comparison of emotional states across different contextual conditions.[106]

BCI passive monitoring

BCI passive monitoring datasets were mainly designed for psychological states, including stress, mental workload, fatigue, and vigilance. Such datasets usually elicited temporal variations in states by adjusting task difficulty or information load. Their defining feature was the use of continuous or ordinal labels to represent state intensity instead of discrete commands or classes. In contrast to control the BCIs that depended on clearly segmented trials, the state-monitoring BCIs focused on long-term EEG signals to track state dynamics for evaluation, early warning, and adaptive modulation. These studies mainly emphasized sustained state tracking and quantification instead of refined decoding of specific cognitive contents.

3.2.3 | Clinical datasets

The clinical category ($n = 116$) primarily consisted of EEG datasets in which patients with specific brain disorders were asked to perform cognitive or resting-state paradigms, enabling comparison of their cognitive or physiological brain states with those of healthy controls. These datasets were commonly used for disease detection, clinical monitoring (e.g., sleep disorders and epilepsy), and characterization of disorder-related brain dysfunction. The datasets on neurological diseases ($n = 41$) and psychiatric disorders ($n = 31$) formed the core of this category, underscoring the extensive application of EEG in probing brain functional abnormalities and neuropsychiatric conditions. To provide a clearer overview of our clinical taxonomy, Figure 5 shows how the clinical datasets are distributed across major disease groups and representative conditions. In the neurological group, common conditions included Alzheimer's disease, Parkinson's disease, brain injury, pain-related conditions, and auditory/cognitive disorders. In the psychiatric group, the most frequent conditions included autism spectrum disorder (ASD), depression, attention-deficit/hyperactivity disorder (ADHD), schizophrenia, PTSD, and related neurodevelopmental or psychosis-spectrum disorders. The epilepsy datasets ($n = 19$) primarily addressed seizure detection, interictal feature characterization, and automated diagnosis. The sleep datasets ($n = 9$) mainly described sleep-stage organization and its associations with disease or functional status. Compared with the EEG datasets developed for cognitive research and BCI, the clinical EEG datasets placed greater emphasis on defined patient groups and disease-specific labels. They were mainly used to characterize clinical populations, improve understanding of disease mechanisms, and support diagnostic and assessment needs in medical practice.

Psychiatric disorders

The Psychiatric EEG datasets mainly addressed abnormal brain function in populations with mental disorders, including ASD, ADHD, and schizophrenia. Such datasets commonly compared patients with healthy controls to delineate atypical neural activity during perceptual, attentional, emotional, or cognitive control tasks. Experimentally, these datasets comprised both resting-state EEG and task paradigms. For instance, dataset #70 integrated motor imitation paradigms with EEG and 3D motion-capture recordings to study EEG–behavior coupling in individuals with ASD during social motor interactions[107]; dataset #216 employed visual paradigms to examine stimulus-locked neural dynamics in adolescents with ADHD[108]; and dataset #133 applied delayed speech production paradigms to explore abnormal neural mechanisms underlying speech preparation and perceptual monitoring in patients with schizophrenia.[109]

Neurological diseases

The EEG datasets on neurological diseases mainly addressed changes in neural activity associated with impaired brain function, including locked-in syndrome, neurodegenerative disorders, brain injury, and disorders of consciousness. In terms of experimental design, these datasets include both resting-state EEG and simple task paradigms related to specific diseases. For example, dataset #703 compared resting-state spectral features between patients with complete locked-in syndrome and healthy controls[110]; dataset #803 used rhythmic motor tasks to analyze alterations in brain activity related to temporal control and motor rhythms in patients with Parkinson's disease[111]; dataset #366 applies inhibitory control tasks to compare oscillatory and functional connectivity features between patients with Alzheimer's disease or frontotemporal dementia and control groups[112]; and dataset #155 employed cognitive control paradigms to evaluate atypical neural oscillations and structure–function coupling in patients with mild traumatic brain injury.[113]

Epilepsy

The epilepsy EEG datasets were primarily used for seizure detection, seizure prediction, and clinical decision support. These data were mostly derived from real-world clinical

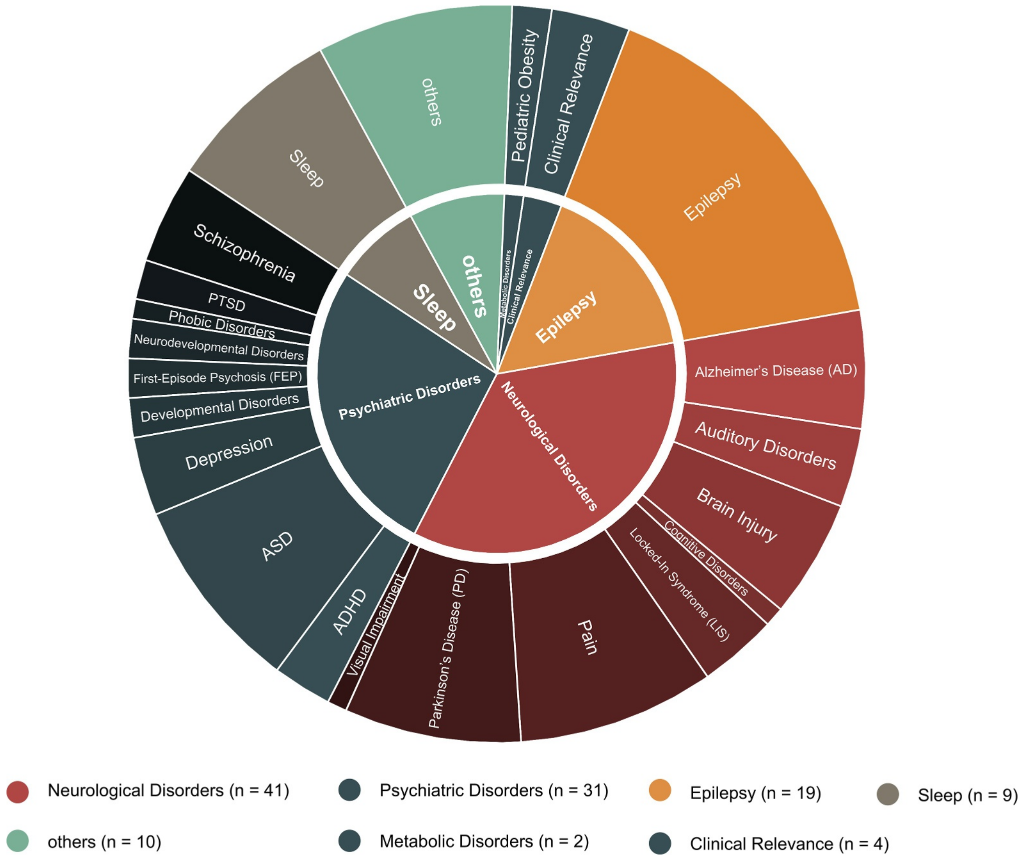

FIGURE 5 Distribution of the clinical EEG datasets across the disease groups and subtypes. ADHD, attention deficit and hyperactive disorder; ASD, autism spectrum disorder; PTSD, post-traumatic stress disorder.

monitoring settings, covering long-duration continuous recordings and multiple seizure stages, with an emphasis on temporal continuity and annotation consistency. Among the public resources, the CHB-MIT and TUH EEG Seizure Corpus had become widely used benchmarks for seizure detection and algorithm evaluation. For example, dataset #776 provided pediatric epilepsy EEG data annotated with high-frequency oscillations.[114] Dataset #926 compiled multichannel EEG data from neonatal epilepsy cases with seizure annotations from multiple experts.[115] Derived from real clinical contexts, these datasets varied widely in recording length and annotation granularity, providing a realistic representation of clinical data complexity and thus serving as appropriate benchmarks for assessing the robustness and practical utility of seizure detection algorithms in real medical settings. Notably, many epilepsy EEG datasets were not centrally hosted but were instead released by research groups or clinical centers via independent websites, resulting in a fragmented data ecosystem.

Others

In addition to the psychiatric, neurological, and epilepsy-related datasets, other clinical EEG resources were organized into several categories with explicit application focuses. One category consisted of sleep-related datasets, exemplified by Sleep-EDF, offering overnight PSG and EEG recordings with manually scored sleep stages and functioning as public benchmarks for studies of sleep architecture and sleep-stage features.[116] Another group encompassed clinically translational datasets not confined to a single disease category; for instance, dataset #393 integrated controlled bed-rest protocols with oddball ERP paradigms to assess the effects of long-term activity restriction on neural function, demonstrating the utility of EEG in quasi-clinical settings.[117] A further category comprised metabolic or special-population datasets; for example, dataset #783 employed task-based EEG to contrast neural activity between obese and non-obese children in cognitive paradigms.[118]

3.2.4 | Naturalistic datasets

The naturalistic datasets category ($n = 55$) mainly examined EEG dynamics in real-world and complex behavioral settings, highlighting continuous interactions between perception, action, and the environment. Overall, perception-driven scenarios predominated, with visual ($n = 14$) and auditory scene ($n = 12$) contexts forming the core, underscoring the dominance of sensory processing studies under naturalistic stimulation. Driving ($n = 8$) and movement ($n = 7$) scenarios also represented important components, emphasizing the applicability of EEG for complex task execution and behavioral monitoring. Relative to laboratory paradigms, the naturalistic datasets exhibited more open task structures and stimulus control and increased noise and individual variability but enabled the study of more ecologically valid cognitive and behavioral states.

Visual scenes
The categorization of visual naturalistic EEG datasets stems from a shift in research focus from highly controlled discrete-stimulus paradigms to neural activity changes in real-world visual contexts. Unlike traditional cognitive tasks, such studies emphasize how the brain processes continuous visual input and dynamically adapts neural activity patterns over time. Accordingly, in this study, the datasets using natural movies, documentaries, or real-world scene videos as stimuli and characterized by long-duration recordings and weak trial structures were classified separately as visual naturalistic datasets. For example, dataset #239 combined movie-viewing tasks with event segmentation annotations to compare neural dynamics across age groups in natural narrative contexts[27]; conversely, dataset #328 used natural documentary content to analyze temporal variations in emotional experience within continuous scenes and their relationship with visual narrative structure.[119] This category reflected a trend in EEG research from within-paradigm effects to dynamic brain function in naturalistic contexts, providing a data foundation for understanding visual processing in real-world environments.

Auditory scenes
The auditory naturalistic EEG datasets investigated EEG dynamics in real auditory environments, emphasizing real-time tracking and processing of sound streams under continuous multi-source conditions. They commonly employed natural speech, narrative materials, or cocktail-party paradigms to assess EEG tracking of target sound streams or semantic content under noisy and continuous semantic input. For instance, dataset #669 integrated natural narrative speech paradigms with continuous modeling to delineate time-locked neural responses to word meaning and contextual information.[28] Additionally, dataset #691 captured participants' continuous tracking of target speech streams in cocktail-party settings to investigate neural mechanisms of speech stream separation and processing.[120]

Driving scenes
The driving scene datasets were built around prolonged, continuous driving tasks as their experimental paradigm. They usually included continuous driving with occasional unexpected events, with synchronized recordings of driving behavior and EEG to examine fatigue buildup, workload variations, and their impacts on decision and control performance. For instance, dataset #754 employed prolonged simulated driving paradigms incorporating random auditory cues and perturbations to relate EEG features to driving performance indices across fatigue progression.[121] In contrast, dataset #547 captured EEG and lane-deviation behaviors under sustained attention in an immersive driving simulator, supporting temporal analyses of driving fatigue and vigilance decrements.[29]

Movements
The movement datasets primarily comprised EEG recordings acquired during continuous natural movements, emphasizing signal variations under real-world movement conditions. These data were often collected during cycling, sustained physical activity, or varying exercise loads to analyze how repetitive movement, sensory feedback, and exercise intensity modulate neural dynamics. For instance, dataset #134 contrasted EEG complexity in resting versus cycling states, demonstrating a general decrease in EEG complexity during periodic movement,[25]while dataset #699 systematically captured EEG data oscillatory dynamics during and following aerobic exercise of varying intensities, delineating the effects of exercise load on cortical rhythms.[26]

In addition to the primary naturalistic scenario categories, a limited number of datasets expand into non-conventional experimental environments and highly open-ended settings. For instance, dataset #534 captured continuous EEG during real MOBA gameplay, reflecting neural dynamics under concurrent complex visual input, rapid decision-making, and sustained motor actions.[122] Dataset #954 acquired EEG data and prefrontal oxygenation measures in natural high-altitude hypoxic settings and, together with Stroop conflict paradigms, contrasted cognitive control and neural dynamics across altitude exposure conditions.[123]

3.2.5 | Neuromodulation datasets

Among the datasets covered in this study, the neuromodulation datasets ($n = 27$) mainly acquired EEG signals concurrently with external stimulation. These datasets spanned multiple stimulation modalities, including transcranial alternating current stimulation (tACS), transcranial magnetic stimulation (TMS), and transcranial direct current stimulatioin (tDCS), as well as other intervention settings (e.g., focused ultrasound stimulation or combined protocols), with EEG data recorded before, during, or after stimulation to examine neural activity differences associated with changes in stimulation conditions.[124–127] In terms of experimental aims, these datasets mainly focused on the relationships between changes in stimulation parameters and EEG rhythms. For example, dataset #161 used concurrent TMS–EEG with stimulation over the frontal eye field to causally probe how top-down signals modulate oscillatory dynamics during visual processing.[12] Dataset #381 applied tACS at different frequencies and stimulation sites during a tonic heat pain paradigm to examine whether oscillatory stimulation modulates pain-related neural and autonomic responses.[128] Conversely, dataset #234 combined anodal tDCS with pharmacological manipulation during a response inhibition task and analyzed task-related theta-band activity to investigate shared modulation mechanisms.[129] Compared with the other categories, these datasets typically involved smaller sample sizes but well-controlled experimental conditions, making them suitable for comparative analyses of stimulation effects and methodological validation.

3.2.6 | Methodological datasets

The methodological datasets ($n = 19$) mainly comprised EEG resources intended for method validation and pipeline evaluation. They included datasets collected under a unified experimental framework across multiple classic paradigms. For example, dataset #439 (ERP CORE) covered several commonly used ERP components, facilitating comparisons of different analysis pipelines and parameter settings.[130] Datasets specifically targeting signal processing problems were also included, such as dataset #386 for evaluating neonatal EEG artifact removal methods,[131] and dataset #277 focused on cardiac field artifact (CFA) correction and single-trial modeling.[132] In addition, some datasets featured methodological explorations in data acquisition, such as **dataset #321**, which used a single amplifier to achieve dual-subject synchronous EEG recordings, providing experimental conditions for multi-agent synchrony analyses. These methodological datasets were intended for testing and benchmarking analysis pipelines, signal processing methods, and experimental organization schemes, rather than for examining specific cognitive or clinical hypotheses.

4 | DISCUSSION

The findings depict the global structure of publicly available EEG data along dimensions such as task categories, application goals, participant cohorts, and recording contexts. These categorizations are intended not only to describe the current landscape but also to answer whether present EEG data span the critical dimensions needed for training general foundation models. From a pretraining perspective, the main limitation in EEG lies not in total data volume but in whether diverse tasks, populations, and scenarios can be systematically incorporated into a unified representation-learning framework. The more than 800 public datasets curated in this study indicate that, although historically dispersed across research themes, they form complementary relationships in cognitive functions, populations, and application goals, providing a practical basis for cross-task representation learning.

In this registry, we define foundation model adaptability in a practical manner: whether users can assemble a pretraining corpus from the registry using clear, searchable metadata. Concretely, the registry supports corpus construction along four dimensions: (i) temporal organization, (ii) channel and montage setup, (iii) label compatibility, (iv) coverage and diversity. For temporal organization, we provide sampling rate, task paradigm, and dataset/publication URL, so users can rapidly confirm the released data format and the way annotations are organized. For channel and montage setup, we present device information, channel count, montage, and modality descriptors to support cross-dataset alignment. For label compatibility, the registry explicitly indicates label availability together with task taxonomy, enabling users to filter datasets for self-supervised or supervised training. For coverage and diversity, we include cohort and acquisition descriptors such as sample size, age, health status, region, device, and release year. Overall, the registry makes foundation model adaptability explicit and reproducible: the corpus can be built by filtering the four dimensions, instead of relying on subjective judgment.

Table 2 provides a qualitative summary of the five dataset categories (excluding the methodological datasets). By relating each category's typical characteristics to pretraining needs (temporal structure, channel/device setup, label structure, and diversity), it connects dataset curation with corpus design and illustrates the repository's adaptability.

4.1 | Resolving category overlap: Illustrative boundary cases

Some datasets naturally overlap across categories. Our rule is simple: we assign one primary category based on the cohort and the main study context, and we keep other relevant information (e.g., paradigm keywords and additional modalities) as metadata. For example, canonical cognitive paradigms can appear in clinically defined cohorts; in such cases, we categorize the dataset as clinical. Dataset #514 uses a frequency-tagging oddball face-processing paradigm in participants with autism and neurotypical conditions, so it is labeled clinical, while oddball/frequency-tagging is retained in the metadata.[133]

A similar overlap occurs in emotion-related datasets. Dataset #945 (SEED-VII) is categorized as BCI because it is explicitly designed for multimodal emotion recognition and decoding, whereas dataset #415 is categorized as cognitive because it focuses on early ERP effects (e.g., P1 modulation) under a controlled task setting.[134]

Finally, naturalistic-looking stimuli alone do not automatically make a dataset naturalistic. Dataset #544 contains simultaneous EEG–fMRI recordings collected during long-form naturalistic stimuli, but it also includes resting-state and a controlled checkerboard paradigms.[135] We

TABLE 2 Qualitative mapping from the dataset categories to the pretraining-relevant features.

Dataset category	Typical characteristics
Cognitive	Time-locked events; condition-level labels
BCI	Decoding targets; structured commands/labels
Naturalistic	Long-form recordings; weak trial boundaries
Clinical	Population heterogeneity; device/site variability
Neuromodulation	Pre/post sessions; repeated measures

categorize it as naturalistic because naturalistic viewing is a major recording context, while the checkerboard task, resting-state, EEG–fMRI setting, and auxiliary modalities are retained as structured metadata.

4.2 | Distributional bias in public EEG resources: Category and platform imbalance

4.2.1 | Category imbalance influenced by cognitive subcategory granularity

Notably, diversity in the current public EEG ecosystem mainly stems from broad coverage of dataset types, not from a balanced number of datasets across types. In our registry, the cognitive datasets comprise the largest group ($n = 504$), while the naturalistic ($n = 55$), neuromodulation ($n = 27$), and methodological ($n = 19$) datasets are much less common. Under this distribution, large-scale pretraining that uses raw category proportions would likely push models toward laboratory cognitive paradigms and their typical signal structure, creating a systematic bias.

Additionally, the numerical dominance of cognitive datasets does not necessarily mean simple redundancy. A key reason is that different categories have different levels of internal granularity in how paradigms and releases are organized. The cognitive category is split into 14 subcategories (e.g., perception [$n = 100$], learning and memory [$n = 90$], language processing [$n = 66$], attention [$n = 60$], emotion processing [$n = 40$], executive function [$n = 35$], and others), as shown in Figure 6. These subcategories often correspond to distinct paradigms and independent dataset releases. In contrast, the other categories are structured more compactly: the BCI datasets are mostly organized into three main types (i.e., control, cognitive decoding, and passive paradigms); the neuromodulation datasets are typically grouped by stimulation modality (e.g., TMS, tDCS, and tACS); the methodological datasets cluster around a small set of technical goals; and the clinical datasets, while diverse in pathology, are usually grouped into a limited number of diagnostic entities. Consequently, registry-level counts partly reflect finer paradigm granularity in cognition rather than straightforward category over-repetition of straightforward categories.

4.2.2 | Platform bias and cross-platform discoverability

Different hosting platforms show clear imbalances in EEG dataset types. In our registry, OSF is dominated by the cognitive datasets (275/353), mostly from traditional laboratory paradigms. OpenNeuro is more balanced between the cognitive and clinical datasets but includes relatively few neuromodulations and methodological datasets. In contrast, Zenodo and Figshare host a higher share of the BCI datasets, especially for control and decoding tasks.

The identified imbalances likely reflect where different research communities tend to share data and how each platform is typically used. OSF is widely adopted in psychology and cognitive neuroscience for project-based sharing, which often matches controlled laboratory studies. OpenNeuro is built around neuroscience sharing workflows and BIDS-style organization, making it easier to release well-documented datasets across domains. Zenodo and Figshare are general-purpose portals frequently used for benchmark releases and engineering-oriented resources, consistent with their larger fraction of BCI decoding and control corpora.

Dataset discovery depends heavily on the hosting platform; searching only one platform can over-sample specific paradigms and increase bias. This is why our registry records the hosting platform for every dataset and links platforms to dataset types, so users can filter by platform and sample across platforms directly from the released table. Additionally, although datasets from different platforms often use different formats, devices, and protocols, we include a concrete example showing that they can still be integrated into a shared representation space for joint modeling when structured metadata are used to guide the process. For this reason, we list platform source information in the registry to support cross-platform selection and integration.

4.2.3 | Broader causes of imbalance

Beyond taxonomy granularity, the shortage of naturalistic datasets also has practical reasons. Naturalistic paradigms have only become common in the last few years, and releasing large naturalistic datasets publicly is often more difficult: the setups are more complex, event markers and annotations are less standardized, and standards documentation and data governance are higher. Similar issues also affect neuromodulation and methodological datasets. These datasets are often built around controlled interventions or specific technical validation goals; therefore, they tend to be smaller and less frequently released at scale.

4.2.4 | Bias risks of proportional sampling and mitigation

If users convert registry-level dataset counts directly into sampling ratios for foundation model pretraining, the training mixture can become biased. In particular, overweighting short, trial-based laboratory paradigms may push the model toward highly structured experimental signals and reduce hurt robustness on continuous clinical recordings, naturalistic settings, or data shifted by interventions. Importantly, our registry does not prescribe any fixed pretraining mixture. Instead, it provides refined metadata so that researchers can build a training corpus that matches their task and generalization goals.

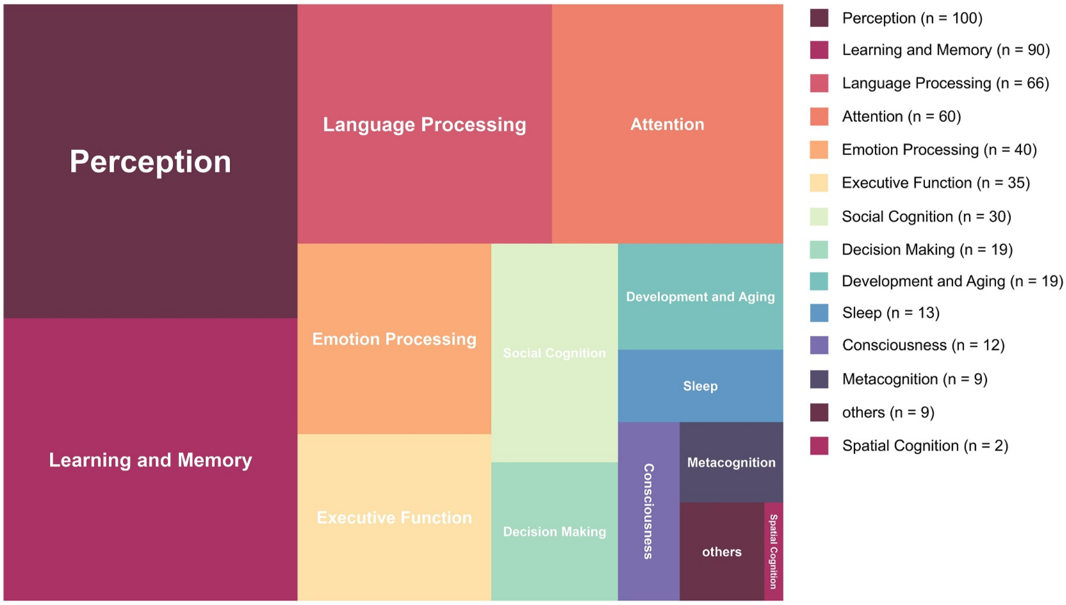

FIGURE 6 Cognitive subcategory distribution with dataset counts.

To reduce imbalance in practice, we suggest simple strategies that can be applied directly to the released table: (i) use domain-aware sampling to control the proportion of major categories, (ii) weight sampling by usable recording duration or window counts rather than by dataset entries, and (iii) utilize the multi-field metadata (e.g., domain/sub-domain, task context, population, device/montage, sampling rate, channel count, modalities, and label availability) for transparent filtering and manual or programmatic rebalancing.

4.3 | Potential omissions and update strategy

4.3.1 | Continuously updated clinical datasets

Even with careful screening, some resources may still be missed. One common reason is that several large clinical EEG databases are continuously updated and are hosted on institutional portals rather than on general-purpose platforms. To ensure transparency about this, we include these resources in the registry and summarize representative examples in Table 3, noting both their scale and whether they are continuously updated. The table covers evolving databases (e.g., TUH-EEG, HBN-EEG, SPaRCNet, and Harvard EEG) as well as widely used benchmark-style static resources (e.g., the European Epilepsy Database and Sleep-EDF Expanded).

For continuously updated databases, new data may appear after our registry snapshot. We therefore explicitly indicate their update status to clarify the time period covered by this release and help users choose datasets consistently for pretraining. TUH-EEG is a prominent example. In our registry, we list its major sub-corpora as separate entries to make them easier to search and filter, and we record its public access status as of January 2026, which is the cutoff for this release.

4.3.2 | Regional platforms and non-English repositories

Regional repositories and non-English data infrastructures can host EEG datasets that are easy to miss if only major global platforms are searched. We conducted targeted searches on regional repositories and added eligible datasets hosted on ScienceDB. We also checked national-level infrastructures (e.g., the National Basic Science Data Center). However, many EEG resources on the platform are not fully open: access is often controlled, submissions are project-based, or redistribution is restricted, which makes large-scale, reproducible retrieval difficult. Therefore, consistent with our inclusion criteria (public access and reproducible retrieval), we did not include these restricted resources in the main registry statistics. We still note them as an important area for future expansion when access conditions become clearer or more open.

4.3.3 | Datasets released only in the referenced articles

We further acknowledge that some EEG datasets are distributed only as journal supplementary files in the referenced articles without independent hosting on a dataset repository or institutional portal. While such releases may contain valuable data, they often lack persistent identifiers,

TABLE 3 Major clinical EEG databases and update types.

Dataset name	Scale	Update nature	Dataset URL
TUH-EEG	Large-scale	Evolving	Dataset Link
HBN-EEG	Large-scale	Evolving	Dataset Link
SPaRCNet	Large-scale	Evolving	Dataset Link
Harvard electroencephalography	Large-scale	Evolving	Dataset Link
European epilepsy database	Benchmark-scale	Static	Dataset Link
Sleep-EDF database expanded	Benchmark-scale	Static	Dataset Link

stable access pathways, and explicit versioning or citation mechanisms, which reduce long-term retrievability and reproducibility. Given that our registry aims to support scalable dataset discovery and reuse for foundation model pretraining, we prioritized resources with independently hosted dataset pages and reproducible access links. We clarified this inclusion boundary in the revised Methods section to explicitly define the scope of the registry.

4.3.4 | Sustainability and update strategy

The registry in this work is meant to be a dynamic resource, not a one-time catalog. Public EEG datasets are growing rapidly across platforms, so we designed the registry to grow with them. We host the release in machine-readable formats with version control, so readers can track updates and cite or reproduce a specific snapshot. We plan to update the registry monthly to add newly released public EEG datasets and keep the collection current.

To keep the registry reliable over time, we will periodically re-check major repositories and verify that datasets are still accessible during each update cycle. When available, we prefer stable identifiers such as DOIs and will document changes in dataset links or availability across versions. Beyond maintenance, the registry is also meant to support gradual standardization. In future releases, we aim to better align with EEG-BIDS, provide harmonized preprocessing examples/pipelines, and improve interoperability with platforms such as OpenNeuro. Over time, we expect the resource to progress from a curated index to a more standardized, interoperable foundation for large-scale EEG research.

4.4 | From dataset indexing to signal standardization: Roadmap and a normalization demonstration

This work provides an EEG dataset registry, not an ETL-style data warehouse that converts raw signals into a single unified format. At this stage, we do not transform all raw EEG data into a fully harmonized corpus. Instead, we provide a structured, multi-field metadata table that indexes public datasets with traceable links, so users can find, filter, and retrieve them reliably.

The registry also records the key information needed for later normalizing—device, channel count and electrode layout, sampling rate, modalities, and annotation details—so it can serve as a practical starting point for data engineering. Building on this, future work can move toward signal-level standardization step by step, such as aligning formats (e.g., EEG-BIDS), mapping channels and montages, normalizing sampling rates, using shared preprocessing protocols, and adding dataset-level quality checks. In this context, the registry serves as an enabling layer for building more interoperable, training-ready EEG resources.

A major challenge for foundation model training is still how to make heterogeneous EEG datasets fit together—datasets recorded with different devices, montages, and protocols do not naturally align. To narrow the gap between registry-level indexing and model training, we provide a lightweight, task-agnostic normalization demonstration as an executable reference. We use four representative datasets (registry IDs #51, #29, #25, and #438) from different repositories and file formats, covering distinct paradigms (RSVP/visual attention, language production, emotion regulation, and SSVEP). They differ in sampling rate (500–1024 Hz), channel density (32–128 channels), and naming conventions, reflecting common real-world variability. As shown in Figure 7, the demo applies a minimal normalization pipeline: fixed-length cropping, average re-referencing, channel alignment to a shared 32-channel 10–20 template (via canonical name matching or geometry-based mapping for EGI systems), resampling to 256 Hz, and 0.5–40 Hz band-pass filtering. The outputs are exported as format-independent tensors with a unified channel order, with event tables preserved when available. The code, configurations, and example outputs are released on GitHub so others can reuse and extend the pipeline.

However, channel normalization involves unavoidable compromises. Mapping high-density systems to a reduced common template can lose spatial detail and blur fine topographic patterns. Geometry-based interpolation and name-based matching also cannot remove all differences due to electrode coverage, density, or head-model assumptions. We therefore treat strict channel matching as a useful baseline, not a one-size-fits-all solution. Beyond explicit channel

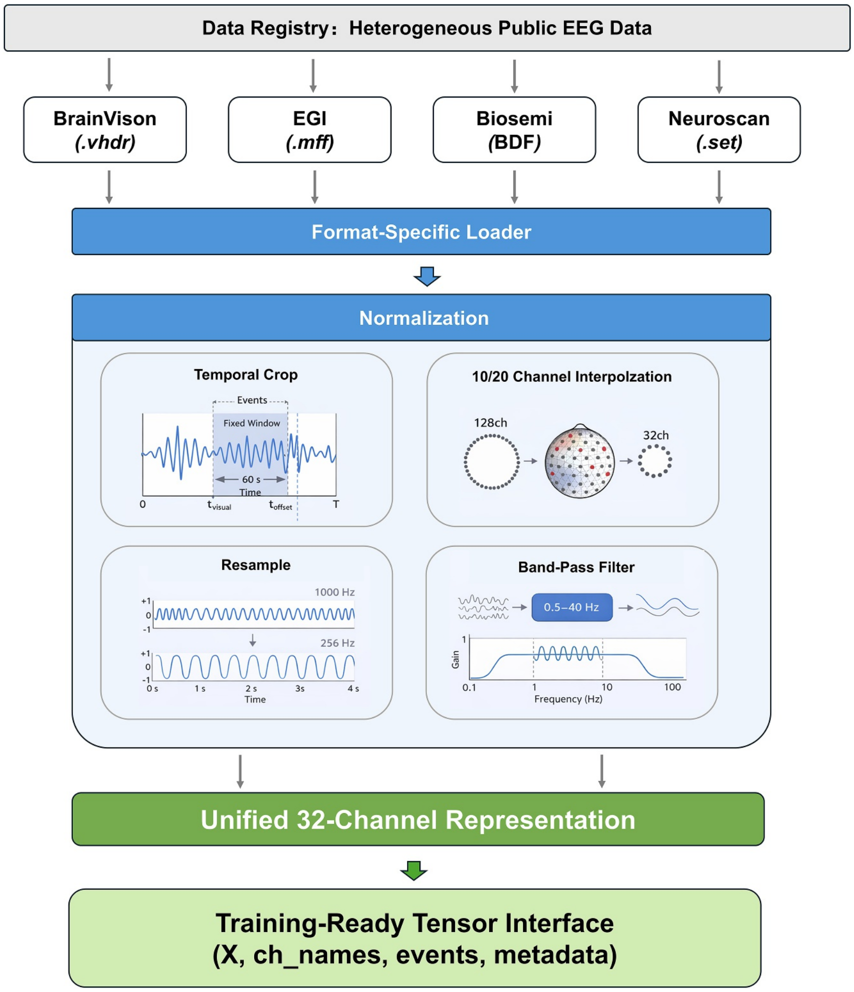

FIGURE 7 EEG normalization workflow demonstration.

alignment, other approaches can reduce reliance on identical montages, such as montage-aware graph models (electrodes as spatial nodes) or spatial filtering projection to move data into a shared anatomical space. These methods do not eliminate montage effects entirely, but they are promising complementary directions for large-scale EEG foundation modeling.

4.5 | Limitations and future updates

The scope of this work is intentionally limited. The registry reflects publicly available EEG datasets within a defined time window; therefore, the reported statistics and classifications may change as new datasets are released and existing resources are updated. Category labels and attribute annotations are taken from original papers and platform descriptions; because terminology and task definitions vary across studies, some boundary cases and inconsistencies are inevitable. In addition, we do not provide quantitative assessments of signal quality or experimental design. Instead, the focus is on discoverability and a structured, machine-readable representation. For these reasons, the registry should be viewed as a practical entry point for dataset discovery and corpus construction rather than a final judgment on the quality of individual datasets. Our next step is systematic standardization. Building on our normalization demonstration, we will progressively improve metadata consistency and expand format/preprocessing alignment for the datasets in the registry, so more entries can be used together in a training-ready manner.

5 | CONCLUSION

Public EEG datasets have long existed as isolated, task-specific, and platform-scattered resources, which substantially limit their reuse for large-scale model development. In this work, we systematically consolidated these fragmented datasets into a unified, machine-readable EEG dataset registry, establishing a structured data space that spans core cognitive functions, clinical populations, applied BCI paradigms, naturalistic environments, and causal neuromodulation settings. Rather than serving as a conventional dataset survey, our registry is explicitly designed as a foundational data infrastructure for large-scale EEG model training. By transforming heterogeneous and scattered EEG resources into a harmonized and accessible data space, this work provides a practical foundation for the development of general-purpose EEG foundation models and enables scalable cross-dataset learning in the EEG domain.

AUTHOR CONTRIBUTIONS

Shengle Shi: Conceptualization; methodology; software; data curation; visualization; writing—original draft; investigation; validation; formal analysis; writing—review and editing. **Yinglu Song**: Investigation; validation. **Yong Wang**: Writing—review and editing; methodology; resources; supervision. **Jiaqing Xiao**: Data curation. **Xinpeng Lin**: Data curation. **Heng Wang**: Data curation. **Zhibin Zhao**: Supervision; resources; methodology; writing—review and editing. **Pengyu Wang**: Methodology; validation; writing—review and editing; supervision. **Zitong Lu**: Conceptualization; methodology; project administration; writing—review and editing; data curation; supervision. **Xiaoli Li**: Funding acquisition; writing—review and editing; methodology; supervision. **He Chen**: Conceptualization; methodology; validation; investigation; funding acquisition; writing—review and editing; project administration; resources; supervision; data curation.

ACKNOWLEDGMENTS

This work was supported by the Strategic International Scientific and Technological Innovation Cooperation Initiative under the National Key R&D Program of China (Grant No. 2025YFE0205100), and by the Science and Technology Projects in Guangzhou (Grant No. 2025A04J4000).

CONFLICT OF INTEREST STATEMENT

Xiaoli Li is the associate editor of the editorial board for Brain-X. The manuscript was handled by another editor and has undergone a rigorous peer-review process. Xiaoli Li was not involved in the journal's review of/or decisions related to this manuscript. The other authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The curated EEG dataset registry constructed in this study is publicly available at Zenodo (https://zenodo.org/records/18815016). The complete multi-field registry is provided in machine-readable formats (.xlsx for Excel and.csv for CSV) with version control. It includes dataset descriptions, standardized metadata fields, traceable access links, and associated references to support reproducible discovery, filtering, and retrieval. Additionally, the demonstration code for data normalization is available on GitHub (https://github.com/sslxd666/eeg-registry-harmonization-demo).

ETHICS STATEMENT

This study used only publicly available datasets and did not involve new recruitment of human participants or animals. Therefore, additional ethical approval was not required.

ORCID

Yong Wang https://orcid.org/0000-0002-1388-7312
Xiaoli Li https://orcid.org/0000-0003-1420-7590
He Chen https://orcid.org/0000-0002-2685-1677

REFERENCES

1. Sanei S, Chambers JA. *EEG Signal Processing*. Wiley; 2008.
2. Dong L, Ke Y, Zhu X, Liu S, Ming D. Long-term cognitive and neurophysiological effects of mental rotation training. *npj Sci Learn*. 2025;10(1):16. https://doi.org/10.1038/s41539-025-00309-2
3. Long NM. The intersection of the retrieval state and internal attention. *Nat Commun*. 2023;14(1):3861. https://doi.org/10.1038/s41467-023-39609-9
4. den Ouden C, Zhou A, Mepani V, Kovács G, Vogels R, Feuerriegel D. Stimulus expectations do not modulate visual event-related potentials in probabilistic cueing designs. *Neuroimage*. 2023;280:120347. https://doi.org/10.1016/j.neuroimage.2023.120347
5. Carter F, DeLuca V, Segaert K, Mazaheri A, Krott A. Functional neural architecture of cognitive control mediates the relationship between individual differences in bilingual experience and behaviour. *Neuroimage*. 2023;273:120085. https://doi.org/10.1016/j.neuroimage.2023.120085
6. Mulders D, Seymour B, Mouraux A, Mancini F. Confidence of probabilistic predictions modulates the cortical response to pain. *Proc Natl Acad Sci U S A*. 2023;120(4):e2212252120. https://doi.org/10.1073/pnas.2212252120
7. Stevenson NJ, Tapani K, Lauronen L, Vanhatalo S. A dataset of neonatal EEG recordings with seizure annotations. *Sci Data*. 2019;6(1):190039. https://doi.org/10.1038/sdata.2019.39
8. Schiff ND. Cognitive motor dissociation following severe brain injuries. *JAMA Neurol*. 2015;72(12):1413-1415. https://doi.org/10.1001/jamaneurol.2015.2899
9. Claassen J, Doyle K, Matory A, et al. Detection of brain activation in unresponsive patients with acute brain injury. *N Engl J Med*. 2019;380(26):2497-2505. https://doi.org/10.1056/NEJMoa1812757
10. Engemann DA, Raimondo F, King JR, et al. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*. 2018;141(11):3179-3192. https://doi.org/10.1093/brain/awy251
11. Carroll EE, Shen Q, Kansara V, et al. Sleep spindles as a predictor of cognitive motor dissociation and recovery of consciousness after acute brain injury. *Nat Med*. 2025;31(5):1578-1585. https://doi.org/10.1038/s41591-025-03578-x
12. Trajkovic J, Veniero D, Hanslmayr S, et al. Top-down and bottom-up interactions rely on nested brain oscillations to shape rhythmic visual attention sampling. *PLoS Biol*. 2025;23(4):e3002688. https://doi.org/10.1371/journal.pbio.3002688
13. Veniero D, Gross J, Morand S, Duecker F, Sack AT, Thut G. Top-down control of visual cortex by the frontal eye fields through oscillatory realignment. *Nat Commun*. 2021;12(1):1757. https://doi.org/10.1038/s41467-021-21979-7
14. Zmeykina E, Mittner M, Paulus W, Turi Z. Weak rTMS-induced electric fields produce neural entrainment in humans. *Sci Rep*. 2020;10(1):11994. https://doi.org/10.1038/s41598-020-68687-8
15. Forenzo D, Zhang Y, Wittenberg GF, He B. Continuous reaching and grasping with a BCI controlled robotic arm in healthy and stroke-affected individuals. *IEEE Trans Neural Syst Rehabil Eng*. 2025;33:3888-3899. https://doi.org/10.1109/TNSRE.2025.3611821
16. Edelman BJ, Meng J, Suma D, et al. Noninvasive neuroimaging enhances continuous neural tracking for robotic device control. *Sci Robot*. 2019;4(31):eaaw6844. https://doi.org/10.1126/scirobotics.aaw6844
17. Ofner P, Schwarz A, Pereira J, Wyss D, Wildburger R, Muller-Putz GR. Attempted arm and hand movements can be decoded from low-frequency EEG from persons with spinal cord injury. *Sci Rep*. 2019;9(1):7134. https://doi.org/10.1038/s41598-019-43594-9

18. Choi GY, Shin JG, Lee JY, et al. EEG dataset for the recognition of different emotions induced in voice-user interaction. *Sci Data*. 2024;11(1):1084. https://doi.org/10.1038/s41597-024-03887-9
19. Luu TP, Nakagome S, He Y, Contreras-Vidal JL. Real-time EEG-based brain-computer interface to a virtual avatar enhances cortical involvement in human treadmill walking. *Sci Rep*. 2017;7(1):8895. https://doi.org/10.1038/s41598-017-09187-0
20. Lee JY, Lee S, Mishra A, et al. Brain-computer interface control with artificial intelligence copilots. *Nat Mach Intell*. 2025;7(9):1510-1523. https://doi.org/10.1038/s42256-025-01090-y
21. Mundanad NA, Zink R, Bertrand A. EEG miniaturization limits for stimulus decoding with EEG sensor networks. *J Neural Eng*. 2021;18(5). https://doi.org/10.1088/1741-2552/ac2629
22. Hinss MF, Jahanpour ES, Somon B, Pluchon L, Dehais F, Roy RN. Open multi-session and multi-task EEG cognitive dataset for passive brain-computer interface applications. *Sci Data*. 2023;10(1):85. https://doi.org/10.1038/s41597-022-01898-y
23. Anders C, Moontaha S, Real S, Arnrich B. Unobtrusive measurement of cognitive load and physiological signals in uncontrolled environments. *Sci Data*. 2024;11(1):1000. https://doi.org/10.1038/s41597-024-03738-7
24. Cannard C, Wahbeh H, Delorme A. Electroencephalography correlates of well-being using a low-cost wearable system. *Front Hum Neurosci*. 2021;15:745135. https://doi.org/10.3389/fnhum.2021.745135
25. Ferre IBS, Corso G, Dos Santos Lima GZ, et al. Cycling reduces the entropy of neuronal activity in the human adult cortex. *PLoS One*. 2024;19(10):e0298703. https://doi.org/10.1371/journal.pone.0298703
26. Ciria LF, Perakakis P, Luque-Casado A, Sanabria D. Physical exercise increases overall brain oscillatory activity but does not influence inhibitory control in young adults. *Neuroimage*. 2018;181:203-210. https://doi.org/10.1016/j.neuroimage.2018.07.009
27. Ghorbani F, Zhou X, Talebi N, et al. Neural connectivity patterns explain why adolescents perceive the world as moving slow. *Commun Biol*. 2024;7(1):759. https://doi.org/10.1038/s42003-024-06439-4
28. Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr Biol*. 2018;28(5):803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080
29. Cao Z, Chuang CH, King JK, Lin CT. Multi-channel EEG recordings during a sustained-attention driving task. *Sci Data*. 2019;6(1):19. https://doi.org/10.1038/s41597-019-0027-4
30. Cavanagh JF, Castellanos J. Identification of canonical neural events during continuous gameplay of an 8-bit style video game. *Neuroimage*. 2016;133:1-13. https://doi.org/10.1016/j.neuroimage.2016.02.075
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016:770-778.
32. Kolesnikov A, Dosovitskiy A, Weissenborn D, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2021.
33. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2020:1877-1901.
34. Devlin J, Chang M.-W, Lee K, Toutanova K. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics; 2019:4171-4186.
35. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. 2015. https://www.wikidata.org/entity/Q104451999
36. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*. Springer International Publishing; 2016:424-432.
37. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. 2021.
38. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022:16000-16009.
39. Xiao Q, Cui Z, Zhang C, et al. BrainOmni: a brain foundation model for unified EEG and MEG signals. 2025.
40. Jiang W.-B, Zhao L.-M, Lu B.-L. Large brain model for learning generic representations with tremendous EEG data in BCI. 2024.
41. Mousavi S, Afghah F, Acharya UR. SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One*. 2019;14(5):e0216456. https://doi.org/10.1371/journal.pone.0216456
42. Wang G, Liu W, He Y, Xu C, Ma L, Li H. EEGPT: pretrained transformer for universal and reliable representatio n of EEG signals. In: *The Thirty-eighth Annual Conference on Neural Information Processing S ystems*; 2024.
43. Yang C, Westover M, Sun J. BIOT: biosignal transformer for cross-data learning in the wild. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2023:78240-78260.
44. Obeid I, Picone J. The temple university hospital EEG data corpus. Data report. *Front Neurosci*. 2016;10. https://doi.org/10.3389/fnins.2016.00196
45. Pernet CR, Appelhoff S, Gorgolewski KJ, et al. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci Data*. 2019;6(1):103. https://doi.org/10.1038/s41597-019-0104-8
46. Markiewicz CJ, Gorgolewski KJ, Feingold F, et al. The OpenNeuro resource for sharing of neuroscience data. *eLife*. 2021;10:e71774. https://doi.org/10.7554/eLife.71774
47. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*. 2000;101(23):e215-e220. https://doi.org/10.1161/01.CIR.101.23.e215
48. Wong S, Simmons A, Rivera-Villicana J, et al. EEG datasets for seizure detection and prediction- A review. *Epilepsia Open*. 2023;8(2):252-267. https://doi.org/10.1002/epi4.12704
49. Tangermann M, Müller KR, Aertsen A, et al. Review of the BCI competition IV. *Front Neurosci*. 2012;6:55. https://doi.org/10.3389/fnins.2012.00055
50. Koelstra S, Mühl C, Soleymani M, et al. DEAP: a database for emotion analysis using physiological signals. *IEEE transactions on affective computing*. 2012;3(1):18-31. https://doi.org/10.1109/t-affc.2011.15
51. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018. https://doi.org/10.1038/sdata.2016.18
52. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-473. https://doi.org/10.7326/M18-0850
53. Luck S. An introduction to the event-related potential technique. 2005.
54. Kikumoto A, Sameshima T, Mayr U. The role of conjunctive representations in stopping actions. *Psychol Sci*. 2022;33(2):325-338. https://doi.org/10.1177/09567976211034505
55. chen y, Wallraven C, Fazli S. An EEG dataset of neural signatures in a competitive two-player game encouraging deceptive behavior. *figshare*. 2024. https://doi.org/10.6084/m9.figshare.24760827.v3
56. Haupt M, Garrett DD, Cichy RM. Data from. *Object recognition in healthy aging (ORHA) - EEG*. 2024. https://doi.org/10.18112/openneuro.ds005363.v1.0.0
57. Bekinschtein TA, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L. Neural signature of the conscious processing of auditory regularities. *Proc Natl Acad Sci USA*. 2009;106(5):1672-1677. https://doi.org/10.1073/pnas.0809667106
58. McFadden KL, Steinmetz SE, Carroll AM, Simon ST, Wallace A, Rojas DC. Test-retest reliability of the 40 Hz EEG auditory steady-state response. *PLoS One*. 2014;9(1):e85748. https://doi.org/10.1371/journal.pone.0085748

59. Zuk NJ, Teoh ES, Lalor EC. EEG-based classification of natural sounds reveals specialized responses to speech and music. *Neuroimage*. 2020;210:116558. https://doi.org/10.1016/j.neuroimage.2020.116558
60. Rutiku R, Fiscone C, Massimini M, Sarasso S. Data from: "auditory oddball comparison (Optimum-1, Learning-oddball, and the local–global paradigm)". 2024. https://doi.org/10.18112/openneuro.ds005429.v1.0.0
61. Bialas O, Schoenwiesner M, Maess B. Data from: "encoding of Sound Source Elevation in Human Cortex". 2022. https://doi.org/10.18112/openneuro.ds004256.v1.0.5
62. Shatek SM, Robinson AK, Grootswagers T, Carlson TA. Capacity for movement is an organisational principle in object representations. *Neuroimage*. 2022;261:119517. https://doi.org/10.1016/j.neuroimage.2022.119517
63. den Ouden C, Zhou A, Mepani V, Kovacs G, Vogels R, Feuerriegel D. Stimulus expectations do not modulate visual event-related potentials in probabilistic cueing designs. *Neuroimage*. 2023;280:120347. https://doi.org/10.1016/j.neuroimage.2023.120347
64. Dolci C, Rashal E, Santandrea E, et al. The dynamics of statistical learning in visual search and its interaction with salience processing: an EEG study. *Neuroimage*. 2024;286:120514. https://doi.org/10.1016/j.neuroimage.2024.120514
65. Kessler R, Skeide MA. Electrophysiological decoding captures the temporal trajectory of face categorization in infants. *Dev Cogn Neurosci*. 2025;75:101601. https://doi.org/10.1016/j.dcn.2025.101601
66. Grootswagers T, Quek GL, Zeng Z, Varlet M. Human infant EEG recordings for 200 object images presented in rapid visual streams. *Sci Data*. 2025;12(1):407. https://doi.org/10.1038/s41597-025-04744-z
67. Khvostov VA, Lukashevich AO, Utochkin IS. Spatially intermixed objects of different categories are parsed automatically. *Sci Rep*. 2021;11(1):377. https://doi.org/10.1038/s41598-020-79828-4
68. Zhao S, Zhao F, Chen Y, et al. Unveiling neurodevelopmental changes in multisensory integration while controlling attention. *J Exp Child Psychol*. 2024;247:106040. https://doi.org/10.1016/j.jecp.2024.106040
69. Gallus V, Hucke CI, Butter K, Ohlmeyer M, van Thriel C. Conceptual processing of natural, complex odours: multisensory effects on behaviour and ERSP. *Brain Res*. 2025;1863:149839. https://doi.org/10.1016/j.brainres.2025.149839
70. Baykan C, Zhu X, Zinchenko A, Muller HJ, Shi Z. Electrophysiological signatures of temporal context in the bisection task. *Exp Brain Res*. 2023;241(8):2081-2096. https://doi.org/10.1007/s00221-023-06670-1
71. Park W, Kim SP, Eid M. Neural coding of vibration intensity. *Front Neurosci*. 2021;15:682113. https://doi.org/10.3389/fnins.2021.682113
72. Ozdemir S, Senturk YD, Unver N, et al. Effects of context changes on memory reactivation. *J Neurosci*. 2024;44(36):e2096232024. https://doi.org/10.1523/JNEUROSCI.2096-23.2024
73. Nasrawi R, Mautner-Rohde M, van Ede F. Memory load influences our preparedness to act on visual representations in working memory without affecting their accessibility. *Prog Neurobiol*. 2025;245:102717. https://doi.org/10.1016/j.pneurobio.2025.102717
74. de Vries IEJ, van Driel J, Olivers CNL. Decoding the status of working memory representations in preparation of visual selection. *Neuroimage*. 2019;191:549-559. https://doi.org/10.1016/j.neuroimage.2019.02.069
75. Tsubomi H, Fukuda K, Kikumoto A, Mayr U, Vogel EK. Task termination triggers spontaneous removal of information from visual working memory. *Psychol Sci*. 2024;35(9):995-1009. https://doi.org/10.1177/09567976241246709
76. van Ede F, Chekroud SR, Stokes MG, Nobre AC. Concurrent visual and motor selection during visual working memory guided action. *Nat Neurosci*. 2019;22(3):477-483. https://doi.org/10.1038/s41593-018-0335-6
77. Wheelock JR, Long NM. Prior memory responses modulate behavior and brain state engagement. *Commun Psychol*. 2024;2(1):121. https://doi.org/10.1038/s44271-024-00165-7
78. Kahana MJ, Rudoler JH, Lohnas LJ, et al. Data from: Penn Electrophysiology of Encoding and Retrieval Study (PEERS). 2023. https://doi.org/10.18112/openneuro.ds004395.v2.0.0
79. Oderbolz C, Stark E, Sauppe S, Meyer M. Concurrent processing of the prosodic hierarchy is supported by cortical entrainment and phase-amplitude coupling. *Cereb Cortex*. 2024;34(12):bhae479. https://doi.org/10.1093/cercor/bhae479
80. Schoknecht P, Yadav H, Vasishth S. Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German. *J Mem Lang*. 2025;141:104599. https://doi.org/10.1016/j.jml.2024.104599
81. Gennari G, Marti S, Palu M, Flo A, Dehaene-Lambertz G. Orthogonal neural codes for speech in the infant brain. *Proc Natl Acad Sci U S A*. 2021;118(31):e2020410118. https://doi.org/10.1073/pnas.2020410118
82. Aguado-Lopez B, Palenciano AF, Penalver JMG, et al. Proactive selective attention across competition contexts. *Cortex*. 2024;176:113-128. https://doi.org/10.1016/j.cortex.2024.04.009
83. van Moorselaar D, Huang C, Theeuwes J. Electrophysiological indices of distractor processing in visual search are shaped by target expectations. *J Cognit Neurosci*. 2023;35(6):1032-1044. https://doi.org/10.1162/jocn_a_01986
84. Herbst SK, Stefanics G, Obleser J. Endogenous modulation of delta phase by expectation-A replication of Stefanics et al., 2010. *Cortex*. 2022;149:226-245. https://doi.org/10.1016/j.cortex.2022.02.001
85. Kosciessa JQ, Lindenberger U, Garrett DD. Thalamocortical excitability modulation guides human perception under uncertainty. *Nat Commun*. 2021;12(1):2430. https://doi.org/10.1038/s41467-021-22511-7
86. Andrillon T, Burns A, Mackay T, Windt J, Tsuchiya N. Predicting lapses of attention with sleep-like slow waves. *Nat Commun*. 2021;12(1):3657. https://doi.org/10.1038/s41467-021-23890-7
87. Ghin F, Stock AK, Beste C. The importance of resource allocation for the interplay between automatic and cognitive control in response inhibition - an EEG source localization study. *Cortex*. 2022;155:202-217. https://doi.org/10.1016/j.cortex.2022.07.004
88. Gyurkovics M, Levita L. Dynamic adjustments of midfrontal control signals in adults and adolescents. *Cereb Cortex*. 2021;31(2):795-808. https://doi.org/10.1093/cercor/bhaa258
89. Clayson PE, Larson MJ. Data from: "Registered Replication Report of ERN/Pe Psychometrics". 2023. https://doi.org/10.18112/openneuro.ds004602.v1.0.1
90. Marly A, Yazdjian A, Soto-Faraco S. The role of conflict processing in multisensory perception: behavioural and electroencephalography evidence. *Philos Trans R Soc Lond B Biol Sci*. 2023;378(1886):20220346. https://doi.org/10.1098/rstb.2022.0346
91. Sun S, Cao R, Rutishauser U, Yu R, Wang S. A uniform human multimodal dataset for emotion perception and judgment. *Sci Data*. 2023;10(1):773. https://doi.org/10.1038/s41597-023-02693-z
92. Daly I, Nicolaou N, Williams D, et al. Data from: "an EEG dataset recorded during affective music listening". 2024. https://doi.org/10.18112/openneuro.ds002721.v10.3
93. Orlowski P, Hobot J, Ruban A, Szczypinski J, Bola M. The relation between naturalistic use of psychedelics and perception of emotional stimuli: an event-related potential study comparing non-users and experienced users of classic psychedelics. *J Psychopharmacol*. 2024;38(1):68-79. https://doi.org/10.1177/02698811231216322
94. Schindler S, Bruchmann M, Krasowski C, Moeck R, Straube T. Charged with a crime: the neuronal signature of processing negatively evaluated faces under different attentional conditions. *Psychol Sci*. 2021;32(8):1311-1324. https://doi.org/10.1177/0956797621996667
95. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. *Clin*

Neurophysiol. 2002;113(6):767-791. https://doi.org/10.1016/s1388-2457(02)00057-3
96. Lotte F, Bougrain L, Cichocki A, et al. A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J Neural Eng*. 2018;15(3):031005. https://doi.org/10.1088/1741-2552/aab2f2
97. Wu D, Lu BL, Hu B, Zeng Z. Affective brain–computer interfaces (aBCIs): a tutorial. *Proc IEEE*. 2023;111(10):1314-1332. https://doi.org/10.1109/JPROC.2023.3277471
98. Wang Y, Chen X, Gao X, Gao S. A benchmark dataset for SSVEP-Based brain-computer interfaces. *IEEE Trans Neural Syst Rehabil Eng*. 2017;25(10):1746-1752. https://doi.org/10.1109/TNSRE.2016.2627556
99. Ding Y, Udompanyawit C, Zhang Y, He B. EEG-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nat Commun*. 2025;16(1):5401. https://doi.org/10.1038/s41467-025-61064-x
100. Meng J, Zhang S, Beyko A, Olsoe J, Baxter B, He B. Author correction: noninvasive electroencephalogram based control of a robotic arm for reach and grasp tasks. *Sci Rep*. 2020;10(1):6627. https://doi.org/10.1038/s41598-020-63070-z
101. Triana-Guzman N, Orjuela-Canon AD, Jutinico AL, Mendoza-Montoya O, Antelis JM. Decoding EEG rhythms offline and online during motor imagery for standing and sitting based on a brain-computer interface. *Front Neuroinf*. 2022;16:961089. https://doi.org/10.3389/fninf.2022.961089
102. Zhao J, Zhang Q, Wang X. YSU-an EEG dataset for studying asynchronous SSVEP-BCI. *figshare*. 2024. https://doi.org/10.6084/m9.figshare.24906300.v3
103. Ke Y, Han Y, Liu P, Ming D. Dataset of binocularly coded steady-state visual evoked potentials recorded with an augmented reality headset. *Sci Data*. 2025;12(1):1338. https://doi.org/10.1038/s41597-025-05696-0
104. Rotaru I, Geirnaert S, Heintz N, Van de Ryck I, Bertrand A, Francart T. What are wereallydecoding? Unveiling biases in EEG-based decoding of the spatial focus of auditory attention. *J Neural Eng*. 2024;21(1):016017. https://doi.org/10.1088/1741-2552/ad2214
105. Zhang Z, Ding X, Bao Y, et al. Chisco: an EEG-based BCI dataset for decoding of imagined speech. *Sci Data*. 2024;11(1):1265. https://doi.org/10.1038/s41597-024-04114-1
106. Xu X, Shen X, Chen X, et al. A multi-context emotional EEG dataset for cross-context emotion decoding. *Sci Data*. 2025;12(1):1142. https://doi.org/10.1038/s41597-025-05349-2
107. Paulo JR, Sousa T, Perdiz J, et al. A multimodal dataset addressing motor function in autism. *Sci Data*. 2025;12(1):959. https://doi.org/10.1038/s41597-025-05313-0
108. Einziger T, Devor T, Ben-Shachar MS, et al. Increased neural variability in adolescents with ADHD symptomatology: evidence from a single-trial EEG study. *Cortex*. 2023;167:25-40. https://doi.org/10.1016/j.cortex.2023.06.007
109. Yang F, Zhu H, Cao X, et al. Impaired motor-to-sensory transformation mediates auditory hallucinations. *PLoS Biol*. 2024;22(10):e3002836. https://doi.org/10.1371/journal.pbio.3002836
110. Maruyama Y, Yoshimura N, Rana A, et al. Electroencephalography of completely locked-in state patients with amyotrophic lateral sclerosis. *Neurosci Res*. 2021;162:45-51. https://doi.org/10.1016/j.neures.2020.01.013
111. Hassall CD, Yan Y, Hunt LT. Data from. *Drum Trainer*. 2022. https://doi.org/10.18112/openneuro.ds004152.v1.1.2
112. Diaz-Rivera MN, Birba A, Fittipaldi S, et al. Multidimensional inhibitory signatures of sentential negation in behavioral variant frontotemporal dementia. *Cereb Cortex*. 2022;33(2):403-420. https://doi.org/10.1093/cercor/bhac074
113. Fox AJ, Matthews N, Qiu Z, Filmer HL, Dux PE. On the lasting impact of mild traumatic brain injury on working memory: behavioural and electrophysiological evidence. *Neuropsychologia*. 2024;204:109005. https://doi.org/10.1016/j.neuropsychologia.2024.109005
114. Cserpan D, Boran E, Rosch R, Biundo SPL, Ramantani G, Sarnthein J. Data from: "dataset of EEG recordings of pediatric patients with epilepsy based o n the 10-20 system". 2021. https://doi.org/10.18112/openneuro.ds003555.v10.1
115. Stevenson N, Tapani K, Lauronen L, Vanhatalo S. Data from: a dataset of neonatal EEG recordings with seizures annotations. 2018. https://doi.org/10.5281/zenodo.2547147
116. Kemp B, Zwinderman AH, Tuk B, Kamphuisen HA, Oberye JJ. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Trans Biomed Eng*. 2000;47(9):1185-1194. https://doi.org/10.1109/10.867928
117. Brauns K, Friedl-Werner A, Gunga HC, Stahn AC. Effects of two months of bed rest and antioxidant supplementation on attentional processing. *Cortex*. 2021;141:81-93. https://doi.org/10.1016/j.cortex.2021.03.026
118. Alatorre-Cruz GC, Downs H, Hagood D, Sorensen ST, Williams DK, Larson-Prior L. Data from: "effect of obesity on arithmetic processing in preteens with high and low math skills. An event-related potentials study". 2022. https://doi.org/10.18112/openneuro.ds004019.v1.0.0
119. Skurka C, Kim N, Eng N, Oliver MB, Awesome A. Emotional flow in environmental messaging. *Media Psychol*. 2024;27(6):869-894. https://doi.org/10.1080/15213269.2023.2297965
120. Das N, Francart T, Bertrand A. Data from: auditory attention detection dataset KULeuven. 2019: 2.0. https://doi.org/10.5281/zenodo.4004271
121. Garcia J, Brooks J, Kerick S, et al. Data from. BCIT Auditory Cueing. 2022. https://doi.org/10.18112/openneuro.ds004105.v1.0.0
122. Li H.-Z, Yang J.-J, Lv Z, et al. Data from: "research data supporting 'EEG recording during playing MOBA game'". 2025. https://doi.org/10.18112/openneuro.ds005520.v1.0.1
123. Ji X, Bao B, Li LZ, et al. EEG and fNIRS datasets based on Stroop task during two weeks of high-altitude exposure in new immigrants. *Sci Data*. 2024;11(1):350. https://doi.org/10.1038/s41597-024-03200-8
124. Hussain SJ, Vollmer MK, Stimely J, et al. Phase-dependent offline enhancement of human motor memory. *Brain Stimul*. 2021;14(4):873-883. https://doi.org/10.1016/j.brs.2021.05.009
125. Schauer G, Chang A, Schwartzman D, et al. Fractionation of parietal function in bistable perception probed with concurrent TMS-EEG. *Sci Data*. 2016;3(1):160065. https://doi.org/10.1038/sdata.2016.65
126. Kosnoff J, Yu K, Liu C, He B. Transcranial focused ultrasound to V5 enhances human visual motion brain-computer interface by modulating feature-based attention. *Nat Commun*. 2024;15(1):4382. https://doi.org/10.1038/s41467-024-48576-8
127. Wischnewski M, Alekseichuk I, Schutter D. Behavioral and electrocortical effects of transcranial alternating current stimulation during advice-guided decision-making. *Neuroimage Rep*. 2021;1(4):100052. https://doi.org/10.1016/j.ynirp.2021.100052
128. May ES, Hohn VD, Nickel MM, et al. Modulating brain rhythms of pain using transcranial alternating current stimulation (tACS) - a Sham-Controlled Study in healthy human participants. *J Pain*. 2021;22(10):1256-1272. https://doi.org/10.1016/j.jpain.2021.03.150
129. Koyun AH, Wendiggensen P, Roessner V, Beste C, Stock AK. Effects of catecholaminergic and transcranial direct current stimulation on response inhibition. *Int J Neuropsychopharmacol*. 2024;27(6):pyae023. https://doi.org/10.1093/ijnp/pyae023
130. Kappenman ES, Farrens JL, Zhang W, Stewart AX, Luck SJ. Erp CORE: an open resource for human event-related potential research. *Neuroimage*. 2021;225:117465. https://doi.org/10.1016/j.neuroimage.2020.117465
131. Arnau S, Sharifian F, Wascher E, Larra MF. Removing the cardiac field artifact from the EEG using neural network regression. *Psychophysiology*. 2023;60(10):e14323. https://doi.org/10.1111/psyp.14323

132. Kumaravel VP, Farella E, Parise E, Buiatti M. NEAR: an artifact removal pipeline for human newborn EEG data. *Dev Cogn Neurosci*. 2022;54:101068. https://doi.org/10.1016/j.dcn.2022.101068
133. Van der Donck S, Hendriks M, Vos S, Op de Beeck H, Boets B. Neural sensitivity to facial identity and facial expression discrimination in adults with autism. *Autism Res*. 2023;16(11):2110-2124. https://doi.org/10.1002/aur.3036
134. Schindler S, Wolf MI, Bruchmann M, Straube T. Fearful face scrambles increase early visual sensory processing in the absence of face information. *Eur J Neurosci*. 2021;53(8):2703-2712. https://doi.org/10.1111/ejn.15111
135. Telesford QK, Gonzalez-Moreira E, Xu T, et al. An open-access dataset of naturalistic viewing using simultaneous EEG-fMRI. *Sci Data*. 2023;10(1):554. https://doi.org/10.1038/s41597-023-02458-8

How to cite this article: Shi S, Song Y, Wang Y, et al. Toward general-purpose foundation models for electroencephalography: a unified data registry. *Brain-X*. 2026;4:e70046. https://doi.org/10.1002/brx2.70046