Human intention inference with a large language model can enhance brain-computer interface control: A proof-of-concept study

Authors

Shue Shiinoki[1], Seitaro Iwama[2], Junichi Ushiba[2*]

[1] Graduate School of Science and Technology, Keio University, Kanagawa, Japan

[2] Department of Biosciences and Informatics, Faculty of Science and Technology, Keio University, Kanagawa, Japan

***Correspondence**

Junichi Ushiba, Ph.D.

Department of Biosciences and Informatics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kouhoku-ku, Yokohama, Kanagawa 223-8522, Japan.

Tel/Fax: +81-45-566-1678.

Email: ushiba@bio.keio.ac.jp

Abstract

Brain-computer interface (BCI) control enables direct communication between the brain and external devices. However, BCI control accuracy with intention inferred from non-invasive modalities is limited, even when using data-driven approaches to tailor neural decoders. In this study, we propose a knowledge-driven framework for inferring human intention, leveraging large language models (LLMs) as an alternative to conventional data-driven approaches. We developed a neural decoder that integrates neural and oculomotor signals with contextual information using an LLM agent. Its feasibility was tested in a real-world BCI task to control a computer application. The LLM-based decoder achieved an average accuracy of 79% for responders (11 out of 20) in inferring the intention to select arbitrary posts in a social networking service. Ablation analyses revealed that integration of contextual information, multimodal signals, and empirical knowledge is critical for decoding accuracy. This study demonstrates the feasibility of a neural decoding framework using an LLM, paving the way for improved performance in BCI-driven external device operation for patients with disability.

Keywords

Brain-Computer Interface; Neural Decoding; Context Inference; Large Language Model; Motor Imagery

Highlights

- Large language models can infer human intent from neural and oculomotor signals integrated with screen context.

- The proposed model performed with higher accuracy compared with data-driven approaches.
- Ablation analyses revealed that integrating contextual information, multi-modal signals, and empirical knowledge is critical for decoding accuracy.

Introduction

A brain-computer interface (BCI) that can bypass human motor pathways to express user volition has been desired for augmented alternative communication[1–5]. To achieve this, a computational model, namely a neural decoder, is employed to infer user intentions embedded in real-time acquired neural signals[6–8]. The decoder plays a critical role in a BCI system, as the decoded results, transformed into device control, need to be accurate enough to reconstruct user intentions. The accuracy from noninvasive modalities has been significantly increased over the decades owing to the development of statistics and machine learning algorithms[9,10].

One of the representative approaches to enhance the BCI performance is deep learning, which is a data-driven representation learning from minimally preprocessed neural signals[8,11,12]. A variety of network structures, such as a recurrent neural network or a convolutional neural network (CNN), have been proposed to extract features from neural signals[13,14] without prior knowledge and assumptions about models. The CNN-based models have demonstrated consistently high performance, leading to the development of derivative models aimed at improving accuracy, generalization across various tasks, and generalization across subjects[8,11,14–19].

Despite astonishing demonstrations using deep learning technology harnessing its high generalization performance and superior accuracy in various fields, such as image processing or sound recognition[20–22], no such breakthroughs have been achieved in neural decoding, especially when using low-fidelity neural measurements represented by scalp electroencephalogram (EEG) as a readout of human cortical activities. It would be because the limited length of experimental time makes it impractical to collect a dataset

covering a vast data space, which is essential to learn a data representation generalizable across datasets, calling for a method to alleviate the fully data-driven nature of existing approaches.

An alternative approach for data-driven model training is to incorporate knowledge-driven signal observation. For instance, in the Bayesian framework, the prior knowledge can be used for initiating the training from a better probability distribution. Further, a more flexible embedding of prior knowledge can be achieved via a large language model (LLM) as a processing block for inference. LLM can be a suitable solution for situations with limited neural training data since it contains neurophysiological knowledge usable for decoding as a supplement to the input signals. Indeed, LLMs have demonstrated human-level, exceptional knowledge-driven reasoning abilities in recommendation systems[23], lifelong learning[24], and autonomous driving tasks[25,26]. For instance, in autonomous driving tasks, LLM systems mimicking human-like decision-making based on common-sense and past experiences have achieved highly generalized autonomous driving through causal reasoning, outperforming existing reinforcement learning-based methods[25,26].

Here, we propose a multimodal decoding system for decoding motor intentions and operation targets from EEG and eye movements based on knowledge-driven reasoning using LLM. We tested its effectiveness in BCI control in a real-world scenario. Participants operated the BCI during the use of social networking service to evaluate the feasibility of the system. The performance of the LLM decoders was compared among models that partially knocked out the input and knowledge, with data-driven models, and with analyses on the discrimination results. We hypothesized that the LLM-based BCI decoder

is achieved by the integration of the observation of multimodal signals, instructions for decoding, and a small number of training data derived from the participants.

Results

Decoder architecture

In this proof-of-concept study, we propose a decoder framework that can integrate EEG and oculomotor signals, contextual information, knowledge, and a few samples of data useful for EEG decoding, which were all integrated and fed into an LLM agent during BCI operation (Fig.1, Supplementary Movie). The LLM agent was instructed to reason out which content shown in the PC display the user selected. All input data were transformed into texts using modules designed for an LLM agent to refer to based on the instruction.

The LLM agent employs the "chain of thought" process for human intention inference, a methodology that has been successfully applied in other fields, such as autonomous driving. In this study, we utilized the ReAct framework[25,27] to facilitate reasoning based on the common knowledge embedded in the LLM, integrating logical reasoning with information extracted from acquired signals, as well as physiological and heuristic knowledge provided as prompts. The React framework addresses the final decision-making problem (i.e., determining where the user is paying attention or whether they intend to press a button) by iteratively selecting an action based on inferred results, observing the outcomes of that action, and then reasoning about those outcomes.

To achieve this, the system leveraged multiple information sources as agent tools (candidate actions), including references to action intention choices, EEG feature calculations, the number of solid viewpoints in each tweet display area, and the spatial

distribution of solid viewpoints. Additionally, environmental data such as tweet and button coordinates, UI information, and physiological knowledge were incorporated as prior information (Fig. 1b). The prompts were generated from text data with integration of those describing EEG decoding knowledge (See Supplementary Information), and those describing the few-shot EEG datasets acquired before the BCI operation, and they were passed to an LLM agent. The model performs the reasoning according to a defined format and finally outputs the operation commands.

The proposed decoder is feasible to infer the user intention

Before the formal evaluation of decoder accuracy, the signal quality of EEG and eye gaze signals was visually inspected. For the EEG signals, the temporal, spatial, and spectral information of EEG used for neural decoding was extracted by spatio-spectral decomposition (SSD) algorithm[28] to visualize time-frequency representations (Supplementary Fig. 1). For the eye gaze signal, we visualized scan path maps indicating sequences of fixations on displays.

First, we asked if the proposed LLM decoder was able to perform neural decoding. We stratified participants based on the responsiveness of scalp EEG signals: the subjects were grouped into responder and non-responder groups. The responders were selected based on SMR-ERD magnitude (Fig. 2a). We evaluated decoding accuracy of action and target selection (i.e, both “push like” or “read” selection and the tweet selection), and that of action detection alone. The results showed that the responder group was superior to the non-responder group in both overall decoding and action decoding alone (Fig. 2b and Supplementary Fig. 2a, *t*-test, $p < 0.05$, Cohen’s $d = 2.86, 2.94$), suggesting that the

proposed decoding approach is feasible for the two groups and dependent on responsiveness of SMR-ERD.

Next, we examined the influence of action decoding accuracy on target decoding accuracy to evaluate the interaction of the two decoding procedures. To this aim, target decoding accuracy was compared between trials in which motor intention was correctly or incorrectly detected. We found that target decoding accuracy was significantly more accurate when motor intention was correctly detected (Fig. 2c, Paired *t*-test, $p < 0.05$, Cohen's $d = 1.88$). Moreover, when target selection was correct, action decoding accuracy was also significantly higher (Supplementary Fig. 2b, Paired *t*-test, $p < 0.05$, Cohen's $d = 1.58$). Therefore, our findings suggest that successful motor intention decoding is associated with higher target decoding accuracy.

Multimodal signal input, prior knowledge, and few-shot information are necessary for effective decoding

In the present study, multimodal data, namely neural and oculomotor movements, were used in an integrated manner for human intention inference. We examined the contribution of each modality to the decoding performance. First, we conducted an ablation study for input modalities (Fig. 3a). The prompt was modified to perform the same inference as the original model. A repeated measures analysis of variance (rmANOVA) revealed that there was a significant decline in the overall decoding accuracy (Main effect: $F = 3.45$, $p < 0.05$; Post-hoc pair-wise *t*-test, EEG (+), Gaze (+) and EEG (+), Gaze (-): $t = 3.3$, $p < 0.05$, Cohen's $d = 1.04$, EEG (+), Gaze (+) and EEG (-), Gaze (+): $t = 2.4$, $p < 0.05$, Cohen's $d = 0.77$), indicating that the decoder with both signals was significantly more accurate than a single decoder for both eye movements and EEG.

Next, we analyzed the impact of two factors on decoding performance: empirical knowledge of EEG decoding, provided in the prompts as essential knowledge for decoding, and few-shot data of EEG features from the first few trials, included as past decoding experience (Fig. 3b). To examine whether knowledge can substitute for experience and whether performance is maintained even with a small amount of few-shot data when knowledge is provided, a two-way rmANOVA was conducted. The independent variables were the presence of empirical knowledge and the amount of few-shot data, while the dependent variable was accuracy. The results showed significant main effects for both factors, but not interaction. Post-hoc tests indicated that when heuristic knowledge was included in the prompt, performance was significantly higher for 0-shot, 1-shot, 5-shot, and 10-shot conditions (Paired *t*-test, $p<0.05$, Cohen's $d = 1.08, 0.98, 0.55, 0.39$, Bonferroni correction for multiple comparisons). Here, the number of "shots" refers to the number of trials provided, with each trial containing four sample data points. Furthermore, 10-shot decoding performed significantly better than both 0-shot and 5-shot conditions (Paired *t*-test, $p < 0.05$, Cohen's $d = 1.84, 0.60$, Bonferroni correction for multiple comparisons).

Comparison with other data-driven decoder architectures

The decoding performance of the knowledge-driven approach was examined in comparison with data-driven methods. Here, we compared the performance with that of CNN and Light Gradient Boosting Machine (Light GBM). The basic type of input for the CNN-based EEG decoder is raw data. However, for the comparison, the CNN model also received SMR-ERD data as input for the LLM in the study.

We compared decoding performance among multiple CNN model and the proposed knowledge-driven model (Fig. 4a). We compared the accuracy of the LLM decoder which uses 10 trials of data from the same subject as few-shot prompts with several versions of EEGNet[17] trained on varying amounts of data: Fine-tuning (10 trials): Trained from scratch using 10 trials of data from the same subject, matching the number of trials used in the LLM decoder. Fine-tuning (50 trials): Trained with 50 trials (i.e., 5 sessions) of data from the same subject. Pretraining + Fine tuning (pretrained + 10 trials): First pre-trained on 500 trials from other subjects, then fine-tuned using 10 trials from the same subject as in the test. The 10-shot LLM decoder showed significantly higher accuracy for the CNN 5 sessions and the CNN 1 session (Paired *t*-test, $p < 0.05$, Cohen's $d = 1.23, 1.09$). This suggests that the performance of the LLM decoder is comparable to that of a CNN decoder with pre-training + finetuning.

To further validate the superiority of the LLM over decision tree-based prediction, we also constructed a model using LightGBM[29] with the same ERD features used as inputs to the LLM. We then compared the decoding accuracy of the 10-shot LLM decoder with that of the LightGBM model trained on data from 5 sessions and 1 session (Fig. 4b). The results showed that the 10-shot LLM was significantly more accurate than the 1-session Light GBM. (Paired *t*-test, $p < 0.05$, Cohen's $d = 1.5$).

Discussion

In the present study, we investigated whether the LLM-based neural decoder architecture is feasible and capable of inferring human movement intention and its target. Through the performance evaluation and ablation analyses, we demonstrate that the proposed LLM-

based neural decoder effectively infers human movement intentions and target selections by integrating neural and oculomotor signals with contextual and empirical knowledge. Specifically, the decoder achieved significantly higher accuracy for responders, with successful motor intention decoding closely linked to improved target selection. Ablation analyses revealed that multimodal signal integration, empirical knowledge, and few-shot learning all significantly contributed to decoding performance, with knowledge-driven inference mitigating the limitations of limited training data. These results support the feasibility of an LLM-driven approach for enhancing BCI control accuracy under a limited amount of data.

The limited amount of available datasets from an individual has critically hindered the establishment of a high-performance decoder for EEG signals. Compared to the fields of image and natural language, the training size is typically in the more than 10 million scale[13,21], with a correspondingly large parameter size and capacity of the model, as well as the data to train it. On the other hand, the deep learning models used in EEG are generally very small due to the small amount of data available for training, making it difficult to capture complex representations. Thus, data-driven algorithms in EEG decoding have attempted to improve performance by compacting parameter sizes and regularization methods[17,18], but still have limited classification performance and generalization capabilities[11,14]. In contrast, the present LLM-based architecture allows researchers to implement decoding techniques and include them in top-down inference assumptions by prompting them. We showed that by leveraging knowledge-driven inference using an LLM, and developed a system architecture that can infer motion intention from less than 50 trials of data. With a small amount of data and a text-based

account of decoding, the model would be beneficial for EEG decoding, where data-driven approaches have been limited by data size.

Previous studies have attempted to augment degrees of freedom by separately using multimodal signals such as gaze movements and EEG. The approach enables simultaneous decoding of interaction targets and behavioral intentions based on gaze and EEG signals, respectively. However, the separate use of these signals could hamper the extraction of information related to movement intention from gaze signals and *vice versa*[30–32]. Specifically, eye movements not only convey information about the object to be manipulated but also reflect behavioral intentions, such as the user's intent to interact with an object when fixating on it or the anticipation of the next task indicated by shifts in gaze[33–36]. The significantly better performance of simultaneous EEG and eye movements input than EEG alone suggests the usefulness of integrated EEG and eye movements decoding. In addition, conventional data-driven methods have the problem that learning information from different modalities in the same model in an integrated manner requires a more complex model due to the no-free-lunch theorem, resulting in a larger data set required. However, LLMs can be easily integrated if a descriptor can be constructed to represent each modal appropriately using prior knowledge and the generic input format of natural language, as demonstrated in the present study.

As demonstrated in the ablation analysis for the prompting for LLM (Fig. 3b), we found that the decoding performance is improved using the few-shot prompting even when physiological or empirical knowledge about EEG decoding is provided as input. If the performance degradation observed with a small amount of few-shot data is mitigated by the inclusion of empirical knowledge, it suggests that the knowledge compensates for the

useful decoding information that the LLM would otherwise extract from the few-shot data. This implies that the empirical knowledge supplements, rather than replaces, information essential for decoding. However, our results indicate that both few-shot prompting and empirical knowledge independently contributed to performance improvements. This suggests that, in addition to the physiological and empirical knowledge provided as text, the LLM also derives useful decoding information from the few-shot data. One possible source of such information is the subject-specific characteristics of EEG signals, particularly in relation to ERD. Previous studies have shown that inter-subject variability in EEG features, such as ERD intensity and variability, constrains the performance of conventional data-driven decoding approaches. Our findings imply that the LLM leverages these subject-specific ERD characteristics from the few-shot data input, thereby enhancing inferential performance.

Although the present study demonstrated that the concept of performing knowledge-driven inference of behavioral intentions, including motor intentions, from a small amount of data using LLM, there are several limitations. First, the inter-/intra-subject generalization of performance was limited by the ability to stably generate SMR-ERDs. Currently, the ability of this inference system to accurately classify behavioral intentions is restricted to evaluation in the best sessions in which subjects with SMR-ERD reactivity above a certain level were able to adequately represent SMR-ERDs throughout. Second, although this inference system inputs all information to the LLM in text form, there are examples of recent LLM research, such as multimodal LLMs that support image input, which have been used in GUI control[37–39] and medical imaging[40]. In the present study, the evaluation was restricted to text input so that the evaluation would not be influenced

by the accuracy with which the LLM extracts features from the input UI and signal images, but image input is beneficial as a means of efficiently inputting information in the same data representation format as when human researchers understand data. Third, the latency of decoding is relatively large, resulting in an approximate 1-minute delay. This delay primarily stems from our reliance on remote LLM inference via APIs, where network overhead and the LLM's processing time significantly contribute to the overall latency. We posit that this issue is technically minor, as it can be resolved by running the LLM locally to eliminate external communication overhead. In addition, ongoing improvements in hardware and software optimizations (e.g., more efficient inference algorithms) will further reduce the response time of LLMs, thus alleviating the current latency challenge.

In summary, we demonstrated that the knowledge-based, LLM-driven neural decoding approach is feasible for BCI control under a limited amount of data. This framework allows the experimenter to design an appropriate prompt as the input for inference, making it possible to use BCI architectures other than motor imagery-based BCIs, such as P300-based speller or steady-state visual-evoked potential BCIs. The proposed approach, based on previous physiological knowledge, would reduce the calibration time without affecting the accuracy, marking a promising step toward practical applications in assistive technologies.

Methods

Participants

Twenty neurologically healthy adults (19 males, 1 female; mean age $\pm$ SD; 24 $\pm$ 1.4 years) were informed of the aim of the experiments and gave informed written consent. The experiment was approved by the local ethics committee of Keio University (IRB

approval number: 2024-009) and performed in accordance with the Declaration of Helsinki. All participants were confirmed as right-handed with the Edinburgh Handedness Inventory[41].

Experimental setup and protocol

EEG was recorded with a 128-channel HydroCel Geodesic Sensor Net from Electrical Geodesics Incorporated (EGI, Eugene, Oregon, USA), with a sampling rate of 1000 Hz. The impedance level of the EEG electrodes was kept below 50 kΩ throughout experiments. In this study, a PC-operated social media application, "X", was used." X" allows users to view short posts, commonly known as tweets, such as informational and news posts from celebrities and friends, and to rate the tweets by 'liking' them. A display showing the screen of an “X” content list was placed in front of the subject. Eye movements were recorded at a sampling rate of 1200 Hz at the bottom of the display with a Tobii Pro Spectrum screen-based eye tracker (Tobii Technology AB, Danderyds, Stockholm, Sweden), which was synchronized with the display to avoid measurement errors. Each subject performed the task of liking a tweet on the “X” GUI by motor imagery while simultaneously measuring EEG and eye movements using the behavioral intention decoding BCI described below.

The experiment consisted of one reference session and five evaluation sessions. After each session, participants completed the Quebec Assistive Technology Satisfaction Evaluation (QUEST) 2.0 questionnaire[42] and a free description questionnaire on their perception of the use of this BCI system to validate its availability. Participants were instructed to rest, read the on-screen content, and press the Like button with the motor image over the time lapse shown in Fig. 1. During the trial, participants were instructed to

perform the task within the same screen without scrolling. To perform the task in a different UI arrangement each trial, scrolling was performed at the end of the trial to update the content to be displayed. The decoder reasoned four times during the trial. The decoder performed reasoning on five seconds of measured data. Reasonings were made twice on “read” and twice on “push like”, for a total of four inferences. The reasoning results were displayed as a GUI during the last “blank” of the trial, and the reasoning results were manipulated with the cursor. The “blank” time was long enough, as inference takes 20-30 seconds.

EEG Processing

EEG analysis was used to input EEG information into the decoding system. The recorded EEG signals were subjected to digital signal processing to remove artifacts. First, a third-order Butterworth bandstop filter (50 Hz) and a third-order Butterworth bandpass filter (3-70 Hz) were applied[43–45]. A spatial filtering technique called spatio-spectral decomposition (SSD) was used on this processed signal, which is capable of extracting neuronal oscillations with spatial patterns that correspond to SM1 activity[28]. The SSD algorithm extracted neural oscillations at the individual alpha frequency (IAF), which reflects the physiological properties of the EEG more effectively than a uniform frequency band[46–48]. IAF was identified using the *specparam* algorithm[49], which parametrizes and separates the periodic and aperiodic components of the EEG signals, and allows calculation of the frequency at which the signal magnitude in the frequency band of interest is maximum.

The input information of EEG to the decoding system was SMR-ERD, which putatively reflects the sensorimotor excitability[50–52]. This is defined as the magnitude of SMR-ERD for motor excitability as follows:

$$ERD(f,t) = \frac{A(f,t) - R(f)}{R(f)}$$

Where A represents the signal power of the EEG signals at time t and f represents the frequency, R and represents the mean signal power of the previous 9 s and recalculated in every 9 s (i.e., at the beginning of relaxation and imagery periods). The power was calculated using a short-time Fourier transform with a 90% overlap of the 1-second sliding window.

Eye movement processing

To input eye movement information into the decoding system, we employed a fixation to represent the position on the screen where the user is gazing[53]. Fixation was extracted by applying an I-VT filter, which extracts gazing and saccades based on eye movement velocity at each time point, with a velocity threshold of 30°/s[54]. Fixations were defined as sequences of gaze points where the velocity remained below the threshold for at least 100 ms. The fixation points were mapped to predefined screen coordinates, and fixation clusters were further quantified per tweet to determine gaze allocation across stimuli.

Decoder Architecture

EEG, eye movements, environmental information in the form of X's UI information, and knowledge of EEG decoding and past EEG data were all integrated and fed into a large language model to reason the user's behavioral intentions. In the signal descriptor, the EEG, eye movements, and UI information were each processed and converted into text

data. The details of each information processing are described below. Prompts were generated from the text data of the signal/environmental information, the EEG decoding knowledge text, and the text data of the measured few-shot EEG in the reference session (see below) and were input into a large language model. The large language model performs the reasoning according to a defined format and finally outputs the operation commands. This was used as input to the application to perform the operations.

ERD time series data for a 5-second window was calculated from the EEG using the online EEG analysis described above. For this time series data, we selected features within the window. As a selection method, we first calculated all the candidate feature values: mean, median, negative 1sd, 2sd, positive 1sd, 2sd, standard deviation, maximum amplitude, location of maximum amplitude, maximum value, minimum value, peak distance mean, zero crossing, number of positive values, number of negative values, and volatility. A gradient boosting model (LightGBM[29]) was then trained to classify classes using these features for the preliminary EEG data, and then the feature importance of the model was calculated, and the eight features that were important were selected as the features to be used. Finally, the mean, median, 1-sd, zero crossing, maximum amplitude value, and location of maximum amplitude, negative percentage, and volatility were selected as features.

The eye movements were used as input by processing the information on the solid viewpoint extracted by the online analysis described earlier. The UI information, like the eye movements, was processed into a form that well represented the user's object of attention by being textualized as the two-dimensional range of existence of the eye movements on the screen and the number of solid viewpoints on each tweet, which was

calculated in combination with the UI information. The text was converted into a form that corresponds to the coordinates of the placement of the tweets and the coordinates of the position of the Like button of each tweet.

Moreover, to facilitate knowledge-driven reasoning, we entered knowledge useful for EEG decoding. The knowledge required for decoding EEG included physiological knowledge to facilitate understanding that SMR-ERD is a biomarker of SM1 excitability during motor imagery, heuristic knowledge in EEG signals observation as to how the signal to be observed is affected by artifacts, and finally few-shot ERD data from EEG measured during the reference session were entered.

Decoder Evaluation

First, significant body motion noise was observed in the data of one of the 20 subjects who performed the experiment, and the quality of the EEG could not be guaranteed, so it was excluded from the analysis. To purely evaluate the performance of the decoder, the remaining 19 subjects were grouped by their SMR-ERD responsiveness, and performance was evaluated for each of these groups. First, a one-sample t-test was conducted in the trial direction for the ERD data of the IAF during each subject's reference session, under the null hypothesis that the mean was zero. Then, subjects whose data for the time interval in which the motor imagery was performed took a significantly non-zero value in the negative direction for more than 50% of the time, as the length of time was considered the High Responder group, and those who did not were considered the Low Responder group.

The performance of the decoder was evaluated based on the accuracy of decoding the presence or absence of the motor intentions “Read” and “Push Like,” as well as the

accuracy of decoding the target tweets that were the target of the action. The decoding performance of the entire behavioral intention, including the target decoding of which tweet was the target of the action, was also evaluated. For the decoding of the entire behavioral intention, we evaluated the self-reported results of which tweets the subject “liked” at the end of the trial as the Ground Truth. In this study, two data sets were obtained for “Push Like” and “Read” in each trial, and since no bias was observed between classes, we evaluated them using overall accuracy.

For data-driven benchmarking, we implemented EEGNet[17]—a compact CNN architecture optimized for EEG-based BCI tasks. The model included three convolutional layers with batch normalization and dropout, and was trained using the Adam optimizer with early stopping based on validation loss. Three training conditions were compared: (1) 10-shot training with subject-specific data, (2) 5-session training from the same subject, and (3) pretraining on 500 trials from other subjects followed by fine-tuning on 10 trials of the target subject. All models used ERD features as input to ensure fairness in comparison with the LLM-based decoder.

We compared the accuracy of the LLM decoder, which includes 10 trials of data in the prompts as few-shot data, and the EEGNet, which is trained with various amounts of data. We prepared a model trained with 10 trials of the same subject's data, which is the same number of trials as the LLM decoder, a model trained with 5 sessions of the same subject's data, which is 50 trials, and a model pre-trained with 500 trials of other subjects' data, and then fine-tuned with 10 trials of the same subject's data. For the training of both models, we employed early stopping, conducted 200 epochs of training, and confirmed that the training converged when the patient's experience was set to 50.

To verify whether decoding by LLM is superior to mere decision tree analysis of each feature, we conducted a Light GBM to classify “Read” and “Push Like”. The model was constructed and evaluated under the same training data conditions as those used for CNN. We also employed early stopping for training and confirmed that the training converged with a boost round of 2000 and an early stopping round of 300.

Statistical Analysis

All data were subjected to parametric tests after testing their normality using the Shapiro-Wilk test. The multilevel datasets were subjected to mixed repeated measures ANOVA and post-hoc pairwise *t*-tests with Bonferroni correction. The sphericity was tested using Mauchly's sphericity test, and Greenhouse–Geisser correction was applied if necessary. Paired datasets were subjected to paired *t*-tests.

Figures

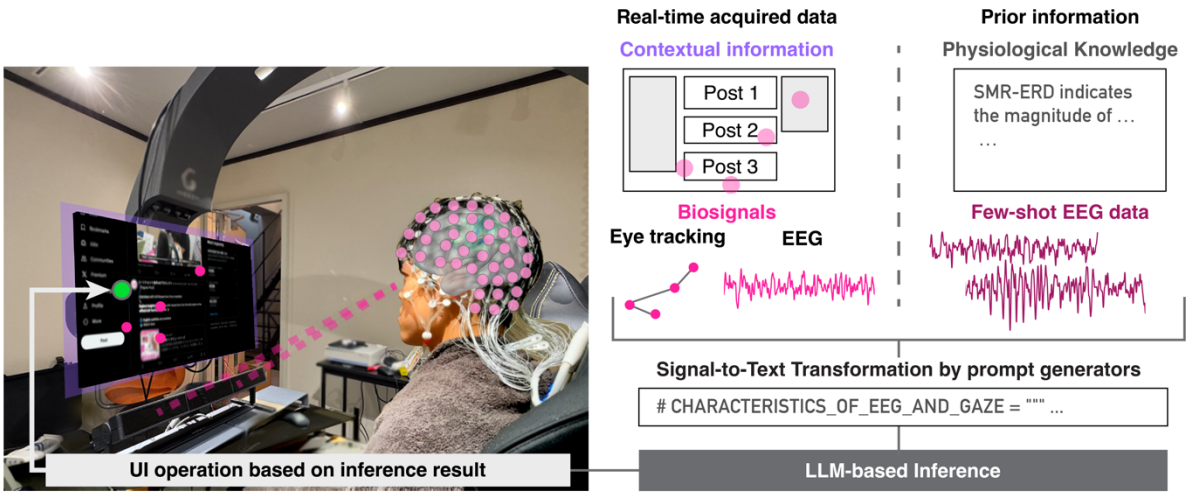

Fig. 1 Schematic of the real-time human intention inference using an LLM agent. The neural and oculomotor signals (i.e., Eye tracker and scalp electroencephalogram (EEG) data) and contextual information (i.e., information about PC display content) were individually processed and converted into text data at the first stage of the processing

pipeline. In parallel, the prior information was converted to text data. The physiological knowledge used for previous EEG-decoding studies, and a few samples of the EEG dataset acquired from the participants, were used for the LLM-based inference. After the inference, the PC cursor moved automatically to the predicted location where the participant aimed.

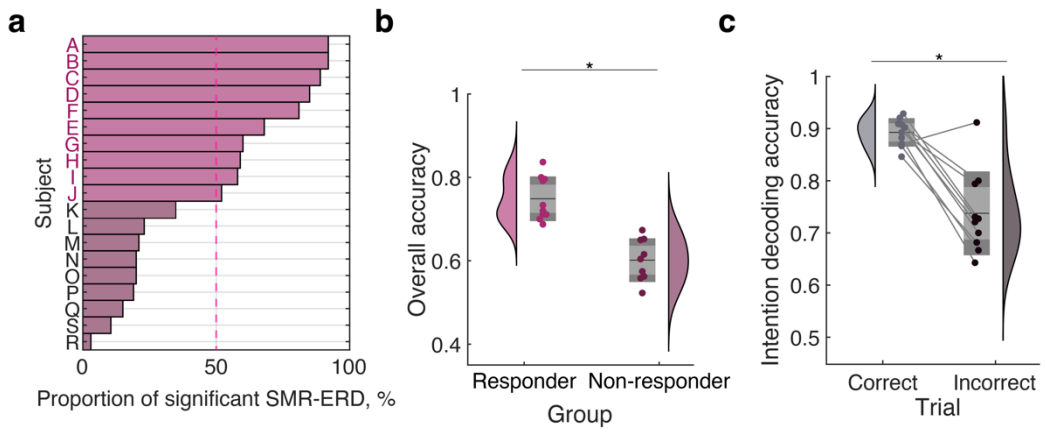

Fig. 2 Evaluation of decoding accuracy during brain-computer interfacing. **a.** Participant-wise proportion of significant SMR-ERD at the first block. A cut-off threshold of 50% accuracy was used to split participants into responder and non-responder groups. Participants included in the responder group were highlighted. **b.** Between-group comparison of overall decoding accuracy for intended action and target. **c.** Comparisons of intention decoding accuracy between trials where the target was correctly or incorrectly estimated.

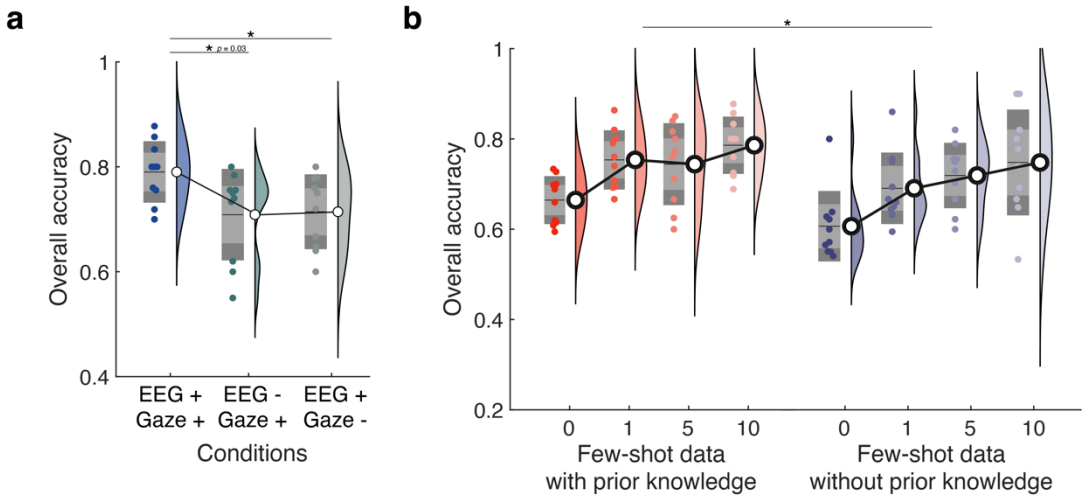

Fig. 3 Multimodal signals and prior knowledge are necessary for inference. **a.** Accuracy of decoding based on multimodal and single-modal data conditions. **b.** Accuracy of decoding with prior knowledge and few-shot data.

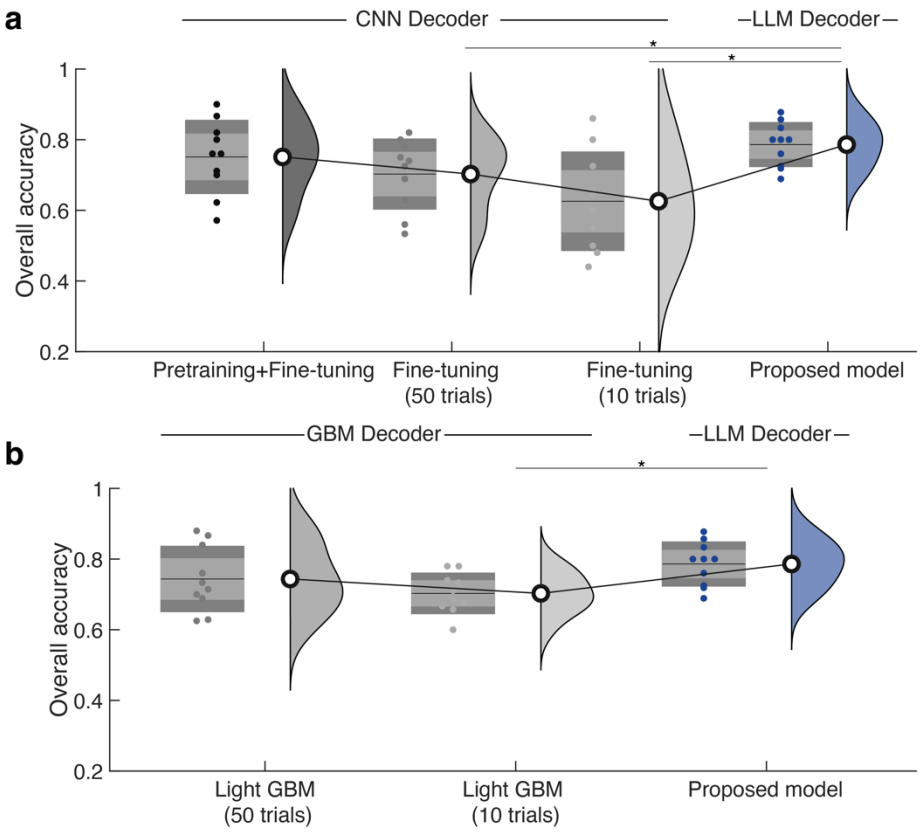

Fig. 4 Comparisons of decoding accuracy with different decoder architectures. a. Decoding accuracy comparison with convolutional neural network (CNN) models. **b.** Decoding accuracy comparison with light gradient boosting machine (GBM) models.

References

1. Edelman, B. J. *et al.* Noninvasive neuroimaging enhances continuous neural tracking for robotic device control. *Science Robotics* **4**, eaaw6844 (2019).
2. Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A. & Chang, E. F. The speech neuroprosthesis. *Nat. Rev. Neurosci.* **25**, 473–492 (2024).
3. Pandarinath, C. *et al.* High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife* **6**, e18554 (2017).
4. 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* **15**, 4382 (2024).

5. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nat Mach Intell* **5**, 1097–1107 (2023).
6. Orsborn, A. L. *et al.* Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. *Neuron* **82**, 1380–1393 (2014).
7. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience* **37**, 435–456 (2014).
8. Schirrmeister, R. T. *et al.* Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping* **38**, 5391–5420 (2017).
9. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F. & Arnaldi, B. A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering* vol. 4 R1–R13 (2007).
10. Lotte, F. *et al.* A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering* **15**, 031005 (2018).
11. Altaheri, H. *et al.* Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review. *Neural Computing and Applications* **35**, 14681–14722 (2023).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
14. Craik, A., He, Y. & Contreras-Vidal, J. L. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering* **16**, 031001 (2019).

15. Tabar, Y. R. & Halici, U. A novel deep learning approach for classification of EEG motor imagery signals. *Journal of Neural Engineering* **14**, 016003 (2017).
16. Zhang, X. *et al.* A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of Neural Engineering* **18**, 031002 (2021).
17. Lawhern, V. J. *et al.* EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering* **15**, 56013–56030 (2018).
18. Dai, G., Zhou, J., Huang, J. & Wang, N. HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification. *Journal of Neural Engineering* **17**, 016025 (2020).
19. Rajwal, S. & Aggarwal, S. Convolutional Neural Network-Based EEG Signal Analysis: A Systematic Review. *Archives of Computational Methods in Engineering* **30**, 3585–3615 (2023).
20. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in 770–778 (2016).
21. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds. Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). doi:10.18653/v1/N19-1423.
22. Brown, T. *et al.* Language Models are Few-Shot Learners. in *Advances in Neural Information Processing Systems* vol. 33 1877–1901 (Curran Associates, Inc., 2020).

23. Yang, S. *et al.* Common Sense Enhanced Knowledge-based Recommendation with Large Language Model. (2024) doi:10.48550/arXiv.2403.18325.
24. Zhu, X. *et al.* Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory. (2023) doi:10.48550/arXiv.2305.17144.
25. Wen, L. *et al.* DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models. (2024) doi:10.48550/arXiv.2309.16292.
26. Fu, D. *et al.* Drive Like a Human: Rethinking Autonomous Driving with Large Language Models. (2023) doi:10.48550/arXiv.2307.07162.
27. Yao, S. *et al.* ReAct: Synergizing Reasoning and Acting in Language Models. Preprint at https://doi.org/10.48550/arXiv.2210.03629 (2023).
28. Nikulin, V. V., Nolte, G. & Curio, G. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. *NeuroImage* **55**, 1528–1535 (2011).
29. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
30. Simola, J., Le Fevre, K., Torniainen, J. & Baccino, T. Affective processing in natural scene viewing: Valence and arousal interactions in eye-fixation-related potentials. *NeuroImage* **106**, 21–33 (2015).
31. Lee, E. C., Woo, J. C., Kim, J. H., Whang, M. & Park, K. R. A brain–computer interface method combined with eye tracking for 3D interaction. *Journal of Neuroscience Methods* **190**, 289–298 (2010).

32. Plöchl, M., Ossandón, J. P. & König, P. Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience* **6**, (2012).

33. Zhu, H. *et al.* Frequency-encoded eye tracking smart contact lens for human–machine interaction. *Nature Communications* **15**, 3588 (2024).

34. Orquin, J. L. & Mueller Loose, S. Attention and choice: A review on eye movements in decision making. *Acta Psychologica* **144**, 190–206 (2013).

35. Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T. & Bunge, S. A. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience* **25**, 69–91 (2017).

36. Skaramagkas, V. *et al.* Review of Eye Tracking Metrics Involved in Emotional and Cognitive Processes. *IEEE Reviews in Biomedical Engineering* **16**, 260–277 (2023).

37. Wang, J. *et al.* Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. (2024) doi:10.48550/arXiv.2401.16158.

38. You, K. *et al.* Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs. (2024) doi:10.48550/arXiv.2404.05719.

39. Nong, S. *et al.* MobileFlow: A Multimodal LLM For Mobile GUI Agent. (2024) doi:10.48550/arXiv.2407.04346.

40. Panagoulias, D. P., Virvou, M. & Tsihrintzis, G. A. Evaluating LLM – Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. (2024) doi:10.48550/arXiv.2402.01730.

41. Oldfield, R. C. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).

42. Demers, L., Monette, M., Lapierre, Y., Arnold, D. L. & Wolfson, C. Reliability, validity, and applicability of the Quebec User Evaluation of Satisfaction with assistive Technology (QUEST 2.0) for adults with multiple sclerosis. *Disabil Rehabil* **24**, 21–30 (2002).
43. He, S., Everest-Phillips, C., Clouter, A., Brown, P. & Tan, H. Neurofeedback-linked suppression of cortical B bursts speeds up movement initiation in healthy motor control: A double-blind sham-controlled study. *Journal of Neuroscience* **40**, 4021–4032 (2020).
44. Kober, S. E., Witte, M., Grinschgl, S., Neuper, C. & Wood, G. Placebo hampers ability to self-regulate brain activity: A double-blind sham-controlled neurofeedback study. *NeuroImage* **181**, 797–806 (2018).
45. Kobler, R. J., Almeida, I., Sburlea, A. I. & Müller-Putz, G. R. Using machine learning to reveal the population vector from EEG signals. *Journal of Neural Engineering* **17**, (2020).
46. Corsi, M. C. *et al.* Functional disconnection of associative cortical areas predicts performance during BCI training. *NeuroImage* **209**, 116500 (2020).
47. Espenhahn, S., Berker, A. O. de, Wijk, B. C. M. van, Rossiter, H. E. & Ward, N. S. Movement-related beta oscillations show high intra-individual reliability. *NeuroImage* **147**, 175–185 (2017).
48. Iwama, S. & Ushiba, J. Rapid-IAF: Rapid Identification of Individual Alpha Frequency in EEG Data Using Sequential Bayesian Estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **32**, 915–922 (2024).

49. Donoghue, T. *et al.* Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience* **23**, 1655–1665 (2020).
50. Ritter, P., Moosmann, M. & Villringer, A. Rolandic alpha and beta EEG rhythms' strengths are inversely related to fMRI-BOLD signal in primary somatosensory and motor cortex. *Human Brain Mapping* **30**, 1168–1187 (2009).
51. Takemi, M., Masakado, Y., Liu, M. & Ushiba, J. Event-related desynchronization reflects downregulation of intracortical inhibition in human primary motor cortex. *Journal of Neurophysiology* **110**, 1158–1166 (2013).
52. Zrenner, C., Desideri, D., Belardinelli, P. & Ziemann, U. Real-time EEG-defined excitability states determine efficacy of TMS-induced plasticity in human motor cortex. *Brain Stimulation* **11**, 374–389 (2018).
53. Carpenter, R. H. S. *Movements of the Eyes*. (Pion Ltd, London, 1977).
54. Munn, S. M., Stefano, L. & Pelz, J. B. Fixation-identification in dynamic scenes: comparing an automated algorithm to manual coding. in *Proceedings of the 5th symposium on Applied perception in graphics and visualization* 33–42 (ACM, Los Angeles California, 2008). doi:10.1145/1394281.1394287.

Conflicts of Interest

J.U. is a founder and representative director of the university startup company, LIFESCAPES Inc., involved in the research, development, and sales of rehabilitation devices, including brain-computer interfaces. SI and He receive a salary from LIFESCAPES Inc., and hold shares in LIFESCAPES Inc. This company does not have any relationships with the device or setup used in the current study. The remaining authors declare no competing interests.

Acknowledgements

This study was supported by JST, PRESTO Grant Number JPMJPR23I1, Japan, and JST Moonshot R&D Grant Number JPMJMS2012, Japan. The authors thank Shoko Tonomoto, Aya Kamiya, and Sayoko Ishii for their assistance.