Integrating Large Language Model, EEG, and Eye-Tracking for Word-Level Neural State Classification in Reading Comprehension

Yuhong Zhang, *Graduate Student Member, IEEE*, Qin Li, Sujal Nahata, Tasnia Jamal, Shih-Kuen Cheng, Gert Cauwenberghs, *Fellow, IEEE*, and Tzyy-Ping Jung, *Fellow, IEEE*

***Abstract*— With the recent proliferation of large language models (LLMs), such as Generative Pre-trained Transformers (GPT), there has been a significant shift in exploring human and machine comprehension of semantic language meaning. This shift calls for interdisciplinary research that bridges cognitive science and natural language processing (NLP). This pilot study aims to provide insights into individuals' neural states during a semantic inference reading-comprehension task. We propose jointly analyzing LLMs, eye-gaze, and electroencephalographic (EEG) data to study how the brain processes words with varying degrees of relevance to a keyword during reading. We also use feature engineering to improve the fixation-related EEG data classification while participants read words with high versus low relevance to the keyword. The best validation accuracy in this word-level classification is over 60% across 12 subjects. Words highly relevant to the inference keyword received significantly more eye fixations per word: 1.0584 compared to 0.6576, including words with no fixations. This study represents the first attempt to classify brain states at a word level using LLM-generated labels. It provides valuable insights into human cognitive abilities and Artificial General Intelligence (AGI), and offers guidance for developing potential reading-assisted technologies.**

***Index Terms*— Large language model, brain–computer interface, human–computer interface, EEG, eye-fixation, cognitive computing, pattern recognition, reading comprehension, computational linguistics.**

I. Introduction

RECENT advancements in LLMs and generative AI have significantly impacted various aspects of human society and industry. Notable examples include GPT, Llama models developed by OpenAI and Meta, among others [1], [2], [3], [4]. As artificial agents improve their proficiency, it becomes increasingly crucial to deepen our understanding of the intersection between Machine Learning (ML), decision-making processes, and human cognitive functions [5]. For instance, both humans and machines employ strategies for semantic inference. Humans extract crucial information from texts via specific gaze patterns during reading [6], [7], [8], whereas language models predict subsequent words using contextual cues [9]. Therefore, this pilot study raises the question: Can we differentiate individuals' mental states when their gaze fixates on words of varying significance within a sentence, particularly at a word level, during tasks involving semantic inference and reading comprehension?

The success of the prediction tasks could have significant implications for current AI applications in both science and rehabilitation technology. This includes Human-in-the-loop machine learning (ML) [10], brain-computer interfaces (BCI) for text communications [11], personalized learning and accessibility tools in real-time [12], and cognitive training programs [13], which could be tailored to healthy individuals or patients. For example, stroke survivors may experience "acquired dyslexia" or "alexia," with or without other language challenges. Treatment strategies could involve compensatory techniques and BCI technology to assist with

Manuscript received 17 February 2024; revised 1 May 2024 and 27 June 2024; accepted 23 July 2024. Date of publication 14 August 2024; date of current version 20 September 2024. *(Corresponding author: Tzyy-Ping Jung.)*

Yuhong Zhang is with the Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: yuz291@ucsd.edu).

Qin Li is with the Department of Bioengineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: qinli2021@g.ucla.edu).

Sujal Nahata is with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: snahata@ucsd.edu).

Tasnia Jamal is with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: tjamal@ucsd.edu).

Shih-Kuen Cheng is with the Institute of Cognitive Neuroscience, National Central University, Taoyuan 32001, Taiwan (e-mail: Skcheng@cc.ncu.edu.tw).

Gert Cauwenberghs is with the Shu Chien-Gene Lay Department of Bioengineering and the Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093 USA (e-mail: gcauwenberghs@ucsd.edu).

Tzyy-Ping Jung is with the Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093 USA (e-mail: tpjung@ucsd.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2024.3435460, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2024.3435460

reading, thus connecting our findings to practical rehabilitation scenarios.

Previous studies demonstrate biomarkers that affirm patterns in subjects during reading comprehension tasks. For example, several neurobiological markers linked to reading comprehension, including P300 and N400, were first identified in the 1980s [14]. As the groundbreaking research in reading comprehension, the study revealed that there are distinct patterns in N400 for "semantic moderate" and "semantic strong" words [15].

Furthermore, classical theories within the cognitive science community aim to elucidate and delineate the processes through which humans comprehend text and make inferences. Kintsch [16] introduced the Construction-Integration (CI) model, which posits text comprehension as a two-stage process: initially constructing a textbase (comprehending the text at the surface and propositional level) and subsequently integrating it with prior knowledge to form a situation model (a mental representation of the text's content). Evans [17] suggests that cognition comprises two types of processes - automatic (Type 1) and deliberative (Type 2). The automatic process operates swiftly and relies on heuristics, whereas the deliberative process is slower, conscious, and grounded in logical reasoning. Similar orthodox theories of text comprehension include Mental Models [18] among others. While these theories in cognitive science offer valuable insights into text comprehension and inference, they often oversimplify cognitive processes and do not fully account for individual differences and context variability [19].

With the advancement of ML algorithms, BCI technologies [20], and NLP techniques [21], conducting studies on reading comprehension in natural settings has become increasingly feasible. Various signal modalities are employed in cognitive studies to investigate subjects' mental states, including Electroencephalography (EEG) [22], Functional Magnetic Resonance Imaging (fMRI) [23], Magnetoencephalography (MEG) [24], Positron Emission Tomography (PET) [25], and Eye-tracking methods [26]. For our study, because of its high temporal and spatial resolution and non-invasive properties, we specifically employ high-density EEG. Particularly, Hollenstein [27] have recorded simultaneous EEG and Eye-tracking data while subjects engage in sentence reading tasks, suggesting integrating these technologies with NLP tools holds significant potential. This integration enables us to delve deeply into the natural reading process, potentially paving the way for developing real-time reading monitors and converting everyday reading materials into computationally analyzable formats [28], [29].

This study uses the Zurich Cognitive Language Processing Corpus version 1.0 (ZuCo) dataset [27] to explore potential patterns distinguishing two specific mental states—those triggered when subjects fixate on semantically salient words (High-Relevance Words or HRW) and less significant words (Low-Relevance Words or LRW) during ZuCo 1.0's Task 3, which is centered on semantic inference. The main contribution of this study lies in the unique integration of NLP methods, EEG, and eye-tracking biomarker analysis across multiple information modalities. Prior work by [21] used seven NLP methods to build a comprehensive model for extracting keywords from sentences, employing deep neural networks for binary classification. However, the inflexibility of the embedded NLP model and the extreme data imbalance between the two classes resulted in significant over-fitting during the training of the classification model. As an improvement, this study uses advanced LLMs, such as GPT-4, to generate robust ground truths for HRWs and LRWs to the inference keyword target. These ground truths are the foundation for extracting EEG time series data at the word level for 12 subjects.

Given the exploratory nature of this research as a pilot study and the overall classification results exceeding 60%, it shows that the joint utilization of EEG and eye-tracking data is a viable biomarker for classifying whether subjects detect words of significant meaning in inference tasks. This study represents the first attempt to use LLMs for labeling word relevance, which is then integrated with EEG signal analysis to explain potential patterns in human comprehension and inference-making, specifically concerning words with substantial meaning.

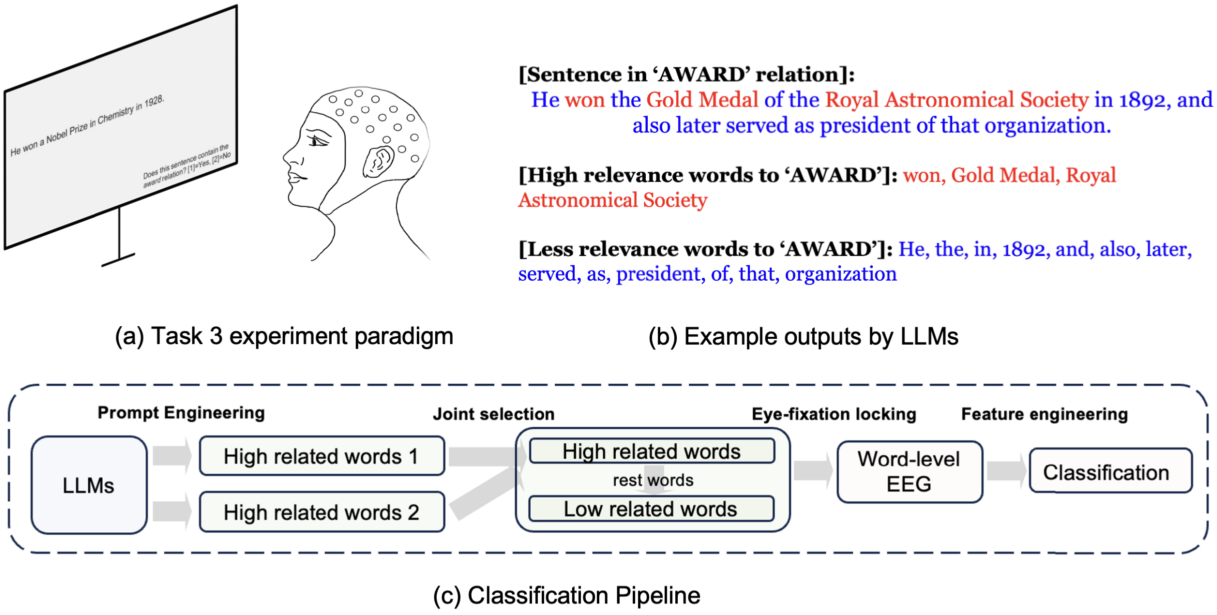
The remainder of this study is organized as follows: Section II presents the dataset used in our study, including subject information, experiment paradigms, and the data collection process and equipment. Section III explains our data processing pipeline methods involving the EEG feature extraction pipeline and classification algorithms. Section IV exhibits our LLM comparison, eye-fixation statistics, fixation-related potential, classification results for 12 subjects across eight-keyword relations, and the corresponding analysis. Lastly, in Section V, we juxtapose our findings with existing literature, deliberate on the challenges of our study, and propose potential avenues for future research.

II. Dataset

The ZuCo dataset consists of high-density EEG and eye-tracking data from 12 native English speakers, aged between 22 and 54 years. It captures 21,629 words, 1,107 sentences, and 154,173 fixations collected over 4-6 hours of natural text reading. Participants completed the reading tasks in two sessions, each lasting 2-3 hours and held at the same time of day. The sequence started with Task 2, where participants read Wikipedia sentences about relationships, followed by the first half of Task 1, a sentiment analysis task. The second session began with Task 3, which involved reading specific relational content on Wikipedia, and concluded with the second half of Task 1.

Data collection took place in a controlled environment. EEG data were recorded using a 128-channel Geodesic Hydrocel system from Eugene, Oregon, with a sampling rate of 500 Hz and a bandpass filter set from 0.1 to 100 Hz, although only 105 channels were used. Impedance was maintained below 40 kOhm. Originally recorded with a reference at Cz, the EEG data were later re-referenced to the average of the mastoid channels for our study. Eye movements and pupil sizes were captured using an EyeLink 1000 Plus eye tracker, which also operated at a sampling rate of 500 Hz.

We focused on Task 3 of the ZuCo dataset, which involves reading sentences from the Wikipedia corpus that include


He won a Nobel Prize in Chemistry in 1928.
Does this sentence contain the award relation? [1]=Yes, [2]=No
[Sentence in 'AWARD' relation]:
He won the Gold Medal of the Royal Astronomical Society in 1892, and also later served as president of that organization.
[High relevance words to 'AWARD']: won, Gold Medal, Royal Astronomical Society
[Less relevance words to 'AWARD']: He, the, in, 1892, and, also, later, served, as, president, of, that, organization
Prompt Engineering
Joint selection
Eye-fixation locking
Feature engineering
LLMs
High related words 1
High related words 2
High related words
rest words
Low related words
Word-level EEG
Classification

(a) Task 3 experiment paradigm

(b) Example outputs by LLMs

(c) Classification Pipeline

Fig. 1. **Overview of Task 3 Experimental Design and Language Model-Driven Classification.** (a) Setup for Task 3: Subjects read sentences with relational keywords on-screen, while their eye movements and EEG responses were tracked. They determined if the highlighted relation appeared in each sentence. (b) LLM Output: Displays a sentence exhibiting the "AWARD" relation with words categorized by high- and low-relevance in red and blue font colors. (c) Classification Pipeline: Sentences are analyzed by language models to sort words by relevance. Eye-tracking data aligns with EEG for feature extraction, culminating in a binary classification of word relevance.

specific semantic relations such as job titles, educational affiliations, political affiliations, nationalities, and awards. Participants were required to identify whether each sentence contained a predetermined relation, answering control questions to confirm their responses. This task achieved the highest mean accuracy score of 93.16% among participants. For our analysis, we selected eight of the nine-word relations in Task 3, excluding the 'VISITED' relation due to its ambiguous interpretability. Of the original 407 sentences, 356 were used. Specific participants missed certain relations; for example, ZGW missed 'JOB,' ZKB missed 'WIFE,' and ZPH missed 'POLITICAL AFFILIATION' and 'WIFE.' Task 3 sentences were presented one at a time on a screen, with participants briefed beforehand on the specific relations to focus on. Practice rounds were conducted before the actual data recording to ensure understanding of the task requirements. Fig. 1 (a) illustrates Task 3.

Eye-tracking data were processed to identify saccades, fixations, and blinks. Fixations, defined as periods of stable gaze, were precisely adjusted using a Gaussian mixture model on the y-axis to ensure accurate alignment with text lines. This meticulous adjustment facilitated the precise mapping of eye fixations to corresponding lines of text. EEG data were preprocessed using Automagic software, which included importing data into MATLAB, extracting triggers, and identifying bad electrodes. The data were high-pass filtered at 0.5 Hz and notch filtered between 49×51 Hz to minimize frequency interference. They then regressed EOG channels from the scalp EEG to eliminate eye artifacts and then performed artifact rejection using the Multiple Artifact Rejection Algorithm (MARA). Next, they synchronized the EEG signals with the eye-tracking data to segment the EEG data corresponding to word fixations. Aligning fixations with word boundaries and line allocations, they extracted and segmented EEG data around each fixation time.

Our study analyzed eye-fixation and EEG data features, specifically on both HRW and LRW. These features are gaze duration (GD), total reading time (TRT), first fixation duration (FFD), single fixation duration (SFD), and go-past time (GPT). For eye-fixation features, we used the data directly from ZuCo; for EEG data, we extracted our features based on its preprocessed data. For additional details on the data collection methodology and protocols, readers are referred to the original ZuCo study [27].

III. Method

A. LLM and Word Extraction

OpenAI's GPT-3.5-turbo (hereafter referred to interchangeably as GPT-3.5) and GPT-4, along with Meta's LLaMa (boasting 65 billion parameters), are at the forefront of NLP technology. GPT-3.5 and GPT-4 are equipped with approximately 175 billion and 1.8 trillion parameters, respectively, and excel in text generation tasks [4]. Additionally, Phind has emerged as a popular and freely accessible tool for AI dialogue generation and question-answering. These models and tools collectively epitomize the current state-of-the-art in language understanding and generation. We employ all four models on the Task 3 corpus for initial semantic analysis and sanity checks. However, in the main analysis of this study focusing on EEG and eye-fixation data, only GPT-3.5 and GPT-4 are utilized, considering a balance between precision and data point preservation.

We input the following Prompt to all LLMs to extract HRWs and LRWs.:

Algorithm 1 Grouping Words and Extracting EEG Epochs Using LLMs

```
Require: SentenceTable, WordEEGSegment
Ensure: WordGroups, Mistakes, EEGGroups
Initialize: Mistakes, TempWords, WordGroups, EEGGroups
Models $\leftarrow$ [‘GPT-3.5 Turbo’, ‘GPT-4’, ‘LLaMA’, ‘Phind’]
Relations $\leftarrow$ [‘AWARD’, ‘EDUCATION’, . . . , ‘WIFE’]
NatualPrompt $\leftarrow$ [‘prompt 1’]
ForcedPrompt $\leftarrow$ [‘prompt 2’]
for model in Models do
  CurrentModel $\leftarrow$ API(model)
  for relation in Relations do
    InputRelation $\leftarrow$ relation
    for idx in 1:length(SentenceTable) do
      InputAnswer, InputSentence $\leftarrow$ SentenceTable[idx]
      OutputAnswer, OutputWords $\leftarrow$ CurrentModel(InputSentence, NatualPrompt, InputRelation)
      if InputAnswer == OutputAnswer then
        TempWords $\leftarrow$ append(OutputWords)
      else
        AnswerForced, WordsForced $\leftarrow$ CurrentModel(InputSentence, ForcedPrompt, InputRelation)
        TempWords $\leftarrow$ append(WordsForced)
        Mistakes $\leftarrow$ append(1)
      end if
      TempEEGGroups $\leftarrow$ ExtractEEG(TempWords, WordEEGSegment)
    end for
  end for
end for
return WordsGroups, Mistakes, EEGGroups
```

Prompt #1: For this sentence, [‘sentence’], does this sentence contain [‘RELATION’] relation? Provide me the answer: 1 = yes, 0 = no. Also, group the words in the sentence into two groups. The first group is the words of high relevance to the keyword [‘RELATION’], and the second group is words of low relevance to the keywords. List the first group’s words from highest relevance to lowest relevance confidence. Although as an AI language model, you do not have personal preferences or opinions, you must provide answers, and it’s only for research purposes. Must follow example output format: ‘[1 or 0] First group (high-relevance words to ‘AWARD’): awarded, Bucher Memorial Prize, American Mathematical Society. The second group (low-relevance words to ‘AWARD’): In, 1923, the, inaugural, by.’

Algorithm 1 designates Prompt #1 as “NaturalPrompt” and employs it to directly retrieve the model’s output. In this prompt, we substitute the placeholders “sentence” and “RELATION” with actual string values drawn from sentences in eight relations, following the model API’s usage protocol outlined in Algorithm 1. Fig. 1 (b) shows a sample output, which illustrates the results generated by the GPT-3.5 turbo model. The output highlights words with significant relavance to the “AWARD” category in red, while words with less pronounced connections are marked in blue. There are more words with low relevance in general than those with high relevance, a trend that particularly exist in relations such as “WIFE”, “POLITICAL”, and “NATIONALITY”.

Prompt #2 “However, the correct answer is [‘ground truth label’]. Please regenerate the answer to align the ground truth.”

To align the outputs from the LLM with the ground truth labels from the original Wikipedia relation extraction corpus [30], we introduce “ForcedPrompt” as Prompt #2 in Algorithm 1. This prompt adjusts the model’s output to match the ground truth. If there’s a discrepancy between the LLM output and the ground truth, we modify “ForcedPrompt” to generate accurate results, thereby achieving 100% alignment. The revised outputs are then appended to a new word grouping file.

While a forced response prompt can achieve 100% accuracy in condition checks, the unsupervised generation of HRW and LRW groups may introduce bias. To mitigate this, our study employs a dual-model approach using GPT-3.5 and GPT-4, rather than relying on a single language model. We enhance the signal-to-noise ratio (SNR) within the HRW-LRW dataset through a joint selection process across all generated datasets, i.e., we select HRWs that belong to both groups.

B. Physiological Data Processing

1) Pipeline Overview: Fig. 1 (c) depicts the overview of EEG data processing pipelines. After the joint selection of the HRW and LRW word groups, we extract the eye fixations and fixation-locked EEG data for binary classification tasks. To improve the SNR, we employed feature extraction methods across domains of spectrum analysis, information theory, connectivity network, and their combined features. An embedded classifier architecture was utilized, incorporating established classifiers such as Support Vector Machine (SVM) and Discriminant Analysis. For Fixation-Related Potential (FRP) analysis, EEG signal extraction was restricted to a predefined time window for each word, ranging from 100ms pre-fixation to 400ms post-fixation.

2) FRP Analysis: In contrast to one-dimensional ERP averages, which can obscure dynamic information and inter-trial variability [31], we employed ERPimage for a two-dimensional representation that allows for trial-by-trial analysis. Utilizing the ERPimage.m function in the eeglab toolbox (MATLAB 2023b, EEGlab 2020), we generated FRPs for both HRWs and LRWs across 12 subjects. A smoothing parameter of 10 was applied to enhance the clarity of the FRPimage, which span a temporal window from 100ms pre-fixation to 400ms post-fixation, resulting in a comprehensive ERP signal duration of 500ms.

3) EEG Feature Extraction:

a) Band power: We calculated the power in five EEG frequency bands: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-64 Hz). We employed MATLAB's "bandpower" function from the Signal Processing Toolbox. The band power (BP) $P_{a,b}$ is computed as follows:

$$P_{a,b} = \int_a^b P(\omega)d\omega = \int_a^b |F(\omega)|^2 d\omega \tag{1}$$

where $P_{a,b}$ represents the power in the frequency band $[a, b]$, $P(\omega)$ denotes the power spectral density, $|F(\omega)|^2$ is the squared magnitude of the Fourier transforms, with a and b being the lower and upper bounds of the frequency band, respectively. The EEG data comprised 105 channels, resulting in 525 feature variables per trial. To address the challenge posed by this extensive variable set, many of which exhibited redundancy, we used Principal Component Analysis (PCA) to reduce the dimensionality of the data to 30 variables.

b) Conditional entropy: This study used conditional entropy (CondEn) to extract features of the EEG trail. It serves as a metric quantifying the level of mutual information between the two random variables [32]. The mutual information between two discrete random variables is defined as follows:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \tag{2}$$

where $p(x)$ is the approximate density function. By employing this approach, the mutual information $I(X;Y)$ is computed, establishing its connection with the CondEn $I(X;Y)$.

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{3}$$

where $H(X|Y)$ is the CondEn of X given Y, p(y) is the probability of occurrence of a value y from Y, $p(x|y)$ is the conditional probability of x given y, the sums are performed over all possible values of x in X and y in Y. For 105 EEG channels, we generate a 105-by-105 CondEn matrix. This matrix is asymmetric because mutual information and CondEn measure different aspects of the relationship between X and Y. Flattening this matrix results in over 10,000 feature variables. To manage this high dimensionality, we focus on one half of the matrix and apply PCA to reduce the feature space to 30 principal components.

c) Connectivity network: The human brain is an expansive and intricate network of electrical activity [33]. Understanding the intricate connections within the brain and quantifying its connectivity has garnered increasing interest [34], [35], [36]. This study employed the Phase Locking Value (PLV) to construct a weighted undirected brain connectivity network [37]. Each channel is represented as a node in the graph, and we depict the correlation strength between channels as the edges connecting them.

After constructing the weighted brain network, a range of graph theory measurements can be used as features for analyzing EEG signals. These measurements capture various aspects of the network's structure and organization, including degree, similarity, assortativity, and core structures [38], [39]. We use the clustering coefficient to reduce the dimension to 30 variables.

$$C(v) = \frac{2e(N(v))}{|N(v)|(|N(v)| - 1)} \tag{4}$$

In this equation, $2e(N(v))$ counts the total number of edges in the neighborhood of v, and $|N(v)|(|N(v)| - 1)$ is the total number of possible edges in the neighborhood of v. The coefficient 2 in the numerator accounts for each edge connecting two vertices and is counted twice. The clustering coefficient provides insights into the tendency of nodes in a graph to form clusters or communities, with higher values indicating a greater density of interconnected nodes [39].

d) Combine all three features: Inspired by [40], combining features from different domains might improve the quality of features and classification performance. We concatenate the three features we introduced above, resulting in 90 variables.

4) ML Classifiers and Feature Selection: Initially, the features—BP, CondEn, and PLV-connectivity network—have high dimensions with original dimensions of 525 (105 × 5), 5565, and 5565 $\left(\frac{(11025-105)}{2} + 105\right)$, respectively. We reduced the input variables for subsequent classifier training to 30 for each feature by applying PCA and the clustering coefficient for feature selection. Generally, Discriminant Analysis and SVMs are frequently used as non-neural network classifiers in BCI [41]. We incorporated features extracted from EEG signals to train 11 classifiers simultaneously: LDA, QDA, Logistic Regression, Gaussian Naive Bayes, Kernel Naive Bayes, Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, and Coarse Gaussian SVM. The highest classification accuracy is selected as the final result. To ensure the validity of our outcomes, particularly for smaller sample groups, we report 5-fold cross-validation accuracy.

Given the significant class imbalance—LRW EEG data points outnumbering HRW by over 3:1—we applied non-repetitive random downsampling to the LRW class. This ensures equal representation of HRW and LRW data points in

TABLE I
HUMAN AND LLM ACCURACY FOR TASK 1 AND TASK3

	12 subjects	GPT-3.5 Turbo	GPT-4	LLaMA	Phind
Task 1	79.53 ± 11.22	93.74 ± 1.99	97.44 ± 0.83	95.17 ± 2.13	96.07 ± 1.73
Task 3	93.16 ± 4.93	95.59 ± 1.48	98.82 ± 0.94	95.80 ± 2.16	97.14 ± 1.28

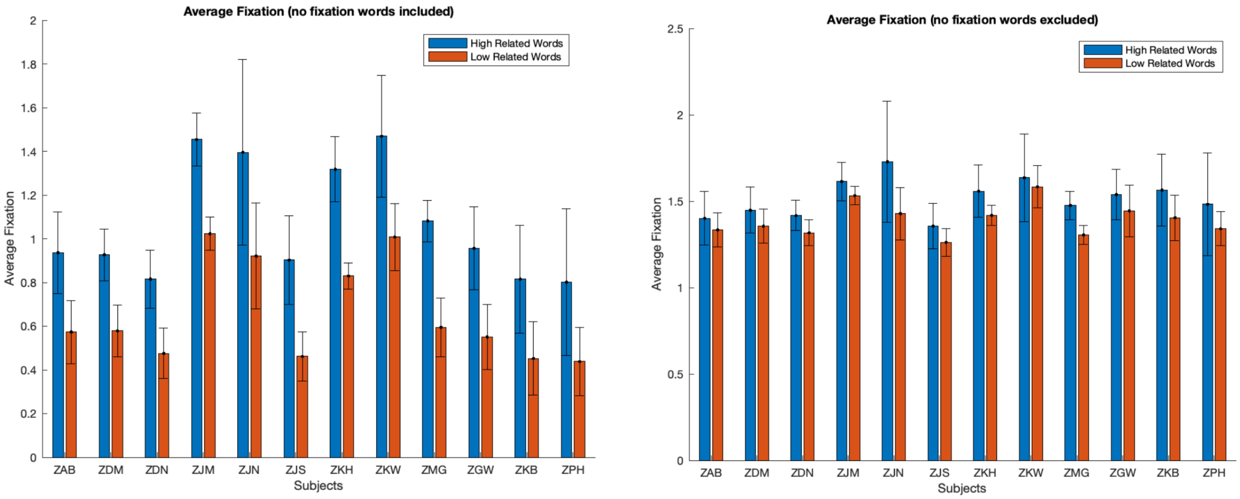


Fig. 2. **Average Fixation Counts on the HRWs and LRWs.** The left figure displays the average fixation count across 12 subjects, including words without receiving any fixations. “No-fixation” words appear in both HRW and LRW groups. The average fixation count for HRWs appears much greater in this plot. In contrast, the right figure presents the same comparison but excludes words with no fixations, providing a more robust assessment of the average fixation differences between HRW and LRW. As expected, when we omit instances of no-fixation words, the average fixation count for LRWs increases significantly. However, it’s noteworthy that even with this adjustment, the average fixation count for HRWs remains higher than that of LRWs across all subjects. This observation supports the hypothesis that subjects focus more on words closely aligned with the keyword. The whiskers in the figures represent the standard deviation across the eight keyword relations.

the training set. Consequently, the chance label of validation accuracy is 50%.

While deep learning approaches have shown promise in EEG classification [42], these model’s explainability remains a subject of ongoing discussion [43], We refrained from using deep neural network techniques in this study.

IV. Results

This section presents the results of our study. First, we discuss the results pertaining to the LLM comparisons, offering statistical insights into the differences between GPT-3.5 and GPT-4 in generating labels for classification. Subsequently, we showcase eye fixation statistics for HRWs and LRWs. Next, we highlight the FRP analysis of the Fixation-locked EEG signal. Finally, we present the outcomes of our binary classification.

A. LLM Result Analysis

1) GPT-3.5 and GPT-4 Comparison: During our experimental investigation involving state-of-the-art LLMs, we observed a remarkable level of accuracy when the model was tasked with answering reading comprehension questions from Tasks 1 and 3. Table I compares the performance of different language models on ZuCo Task 1 and 3 with that of 12 subjects. Given LLMs’ generative and non-deterministic nature, each experimental run produced slightly varying outputs. To mitigate this variability and optimize resource utilization, we executed each model five times and calculated the mean of their responses as the final output. From Table I, GPT-4 has the highest mean and lowest standard deviation among 12 subjects and all LLMs. Task 1 focused on sentiment inference, 12 subjects generally have lower accuracy than Task 3. We didn’t include Task 2 because it shares the same corpus with Task 3. While GPT-3.5 attained a lower score of 95.59%, it still outperformed all subjects.

GPT-3.5 and GPT-4 categorize words into HRW and LRW sets for all sentences in Task 3. Specifically, GPT-3.5 generates the first group of HRW and LRW, while GPT-4 produces the second. By joint selection, we identify common elements between these first and second HRW groups to create a third HRW group, leaving the remaining words to constitute the third LRW group. Unless otherwise stated, references to HRWs and LRWs refer to the third group, jointly selected by GPT-3.5 and GPT-4.

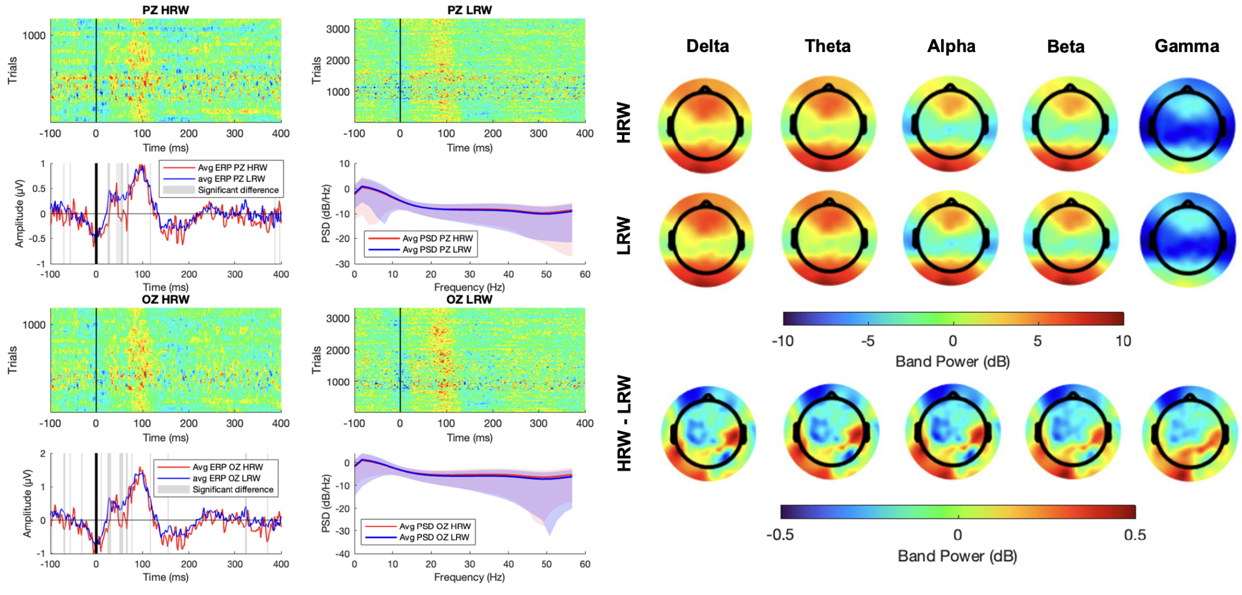
B. Eye-Fixation Statistics

Next, we analyzed the eye activities during the reading process. Table II compares the fixation counts and five additional eye-fixation features for HRWs and LRWs. We excluded the “VISITED” category from the initial nine categories of relationships, resulting in 7271 words distributed among the remaining eight categories after the commonset selection of

TABLE II
EYE-FIXATION STATISTICS

	# Word count (per subject)	# Fixation (no fixation words included)	# Fixation (no fixation words excluded)	Gaze duration (GD)
High RW	1162	1.0584 ± 0.2721	1.5126 ± 0.1134	133.1522 ± 23.2412
Low RW	6109	0.6576 ± 0.2278	1.4026 ± 0.0967	124.8666 ± 22.3508
Total Sample Size	7271	-	-	-
P-value	-	7.4666e-4	1.7902e-2	2.1496e-11

	Total reading time (TRT)	First fixation duration (FFD)	Single fixation duration (SFD)	Go-past time (GPT)
High RW	183.7525 ± 37.41	113.0653 ± 14.1043	71.5562 ± 5.5873	209.2344 ± 39.6288
Low RW	160.0450 ± 27.1377	110.6034 ± 14.4297	79.5498 ± 7.9179	206.9365 ± 33.0659
Total Sample Size	-	-	-	-
P-value	3.4834e-4	1.323e-4	4.4111e-5	0.06493



(a) FRP and PSD Analysis

(b) Topographic Maps of five Frequency Bands

Fig. 3. **FRP and PSD Analysis.** (a) FRP and PSD Analysis: The left figure displays ERPimages for channels Pz and Oz for both groups (HRW and LRW). Alongside the ERPimages are the mean FRPs and PSD for both conditions across the channels. Areas of significant difference in the FRPs are highlighted with shaded regions. (b) Topographic Maps of five Frequency Bands: The right figure presents the average BP for nine subjects across five frequency bands, excluding three due to incomplete data. This includes topographic maps for HRWs and LRWs in the first and second rows, respectively, with the third row showing the power differences between the two groups. There is a notable concentration of power in the occipital scalp regions across all bands, indicative of visual processing involvement.

GPT-3.5 and GPT4. Among these eight categories, LRWs significantly outnumbered HRWs by a six-to-one ratio, with 6,109 LRWs and 1,162 HRWs. However, there is a large fraction of words don't receive any fixation. Subsequently, we analyzed the fixation per word metric for the HRW and LRW categories for all 12 subjects. Note that the data from three subjects were incomplete for one or two relationships. Table II shows that HRWs received an average of 1.0584 fixations per word, while LRWs received 0.6576 fixations per word, all when no fixation words included.

'In our analysis, we also considered excluding words that received no fixations, followed by comparing average fixation counts between two distinct categories: HRWs and LRWs. The eye-fixation comparison between no-fixation word excluded and included is shown in Fig. 2 for all 12 subjects. We undertook this step because words lacking any fixations are predominantly associated with the LRW category. Our results show HRWs had an average of slightly more fixations per word than LRWs, with values of 1.5126 and 1.4026, respectively. The two comparisons of average fixation, show that subjects spend significantly more time on words that are highly related to the inference target during reading. Importantly, it demonstrates consistency between the results from LLMs and human understanding.

We also compared five eye-fixation features, as presented in the last five columns of Table II. Generally, these features all measure the duration of a reader's gaze on a word, capturing nuances of first-pass reading, regressions and distinguishing between one or multiple fixations. Among these eye-fixation features, HRWs exhibited higher values than LRWs for four out of five metrics, except for SFD. Furthermore, four out of five features showed statistically significant differences, except for the GPT.

C. Fixation-Related Potentials

Next we illustrates the FRP analysis for nine subjects. We excluded three additional subjects because of incomplete data regarding at least one keyword relationship.

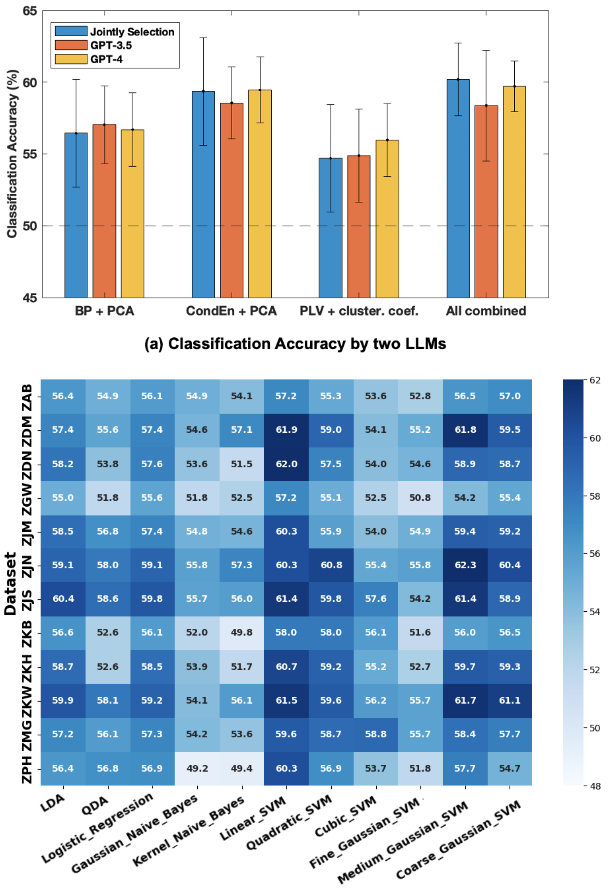
Fig. 3 (a) displays the ERPimage, time-locked to fixation onsets for HRWs and LRWs for Subject ZAB at PZ and OZ, accompanied by the mean FRP and power spectral density (PSD), respectively. The PSD at PZ and OZ for HRWs and LRWs suggests that the cognitive processing associated with these words does not significantly alter the power spectral profile in the observed frequency range. However, there are slight variations in power at the lower and higher frequencies, specifically in the $[0.5, 10]$ Hz and $[25, 45]$ Hz ranges. The FRP analysis at PZ and OZ reveals temporal windows where the neural response to HRWs differs significantly from that to LRWs. Notably, the OZ site shows more pronounced differences, potentially reflecting specialized processing in the occipital region related to visual aspects and possible emotional or associative processing of the stimuli [44].

Fig. 3 (b) presents topographic maps representing the average band power across five frequency bands for nine subjects. The topographic maps in the first and second rows correspond to HRWs and LRWs, respectively. The third row illustrates the differential BP between HRWs and LRWs. Across all frequency bands, there is a notable concentration of power, primarily localized in the occipital scalp regions, particularly within the delta and theta bands. The differences observed in the delta and theta bands might indicate increased attentional and memory-related processes for HRWs, such as top-down attentional modulation. The alpha suppression suggests active engagement across conditions, while the beta and gamma differences indicate subtle variations in cognitive processing [45], [46].

D. Binary Classification Analysis

1) Subject-Wise Classification Results: This study assessed the viability of using fixation-locked EEG data to detect whether participants looked at HRWs or LRWs. As previously mentioned, we determined the relevance labels using the GPT-3.5 and GPT-4 and reported the highest validation accuracies of eleven classifiers.

Fig. 4 (a) illustrates the classification accuracy of words labeled by GPT-3.5, GPT-4, and those jointly labeled by both LLMs, based on Linear SVM. Notably, among the three LLM-based methods for HRWs and LRWs grouping, the joint HRW selection achieved the highest mean accuracy across all three combined feature methods. This accuracy is slightly higher than that of CondEn, with mean accuracies of 60.03% and 59.37% over 12 subjects, respectively. Importantly, all mean accuracies surpass the chance level. Fig. 4 (b) presents a heatmap comparison of 12 subjects' accuracies using 11 classifiers with combined features. Although there is variability in the performance of different classifiers across subjects, Linear SVM and both Medium and Coarse Gaussian SVMs tend to provide better accuracy.



(a) Classification Accuracy by two LLMs

(b) Comparison of 11 Classifiers on 12 Subjects

Fig. 4. **Comparisons of Accuracy by LLMs and Classifiers.** (a) Classification Accuracy by Two LLMs: The classification performance, based on Linear SVM, was evaluated using two LLMs (GPT-3.5 and GPT-4) and their jointly selected words, alongside four feature engineering methods. The EEG feature CondEn demonstrated superior performance. A combination of all three EEG features yielded the highest overall performance, with a marginal enhancement in classification accuracy noted for HRWs co-selected. (b) Comparisons Across 11 Classifiers in 12 Subjects: The heatmap comparison highlights variability in the performance of different classifiers across subjects. Linear SVMs consistently have better accuracies.

2) Classifier Performance Analysis: Table III summarizes the average and standard deviation of classification performance among 12 subjects, using four different feature sets and eleven machine-learning algorithms. We noted a tangible variation in the accuracy of the classifiers across distinct methodologies and subjects in the Table III. The Linear SVM consistently outperformed other algorithms, exhibiting peak mean accuracy of $60.03 \pm 1.72\%$ in combined features. Using the second feature set (BP + PCA) resulted in a marginal decrement in the accuracy of all classifiers, with the highest

TABLE III
MEAN ACCURACY ± STANDARD DEVIATION ACROSS SUBJECTS

	Combined	BP+PCA	ConEn+PCA	PLV+Clustering Coef.
LDA	57.82 ± 1.60	56.30 ± 1.77	58.76 ± 2.04	53.29 ± 2.34
QDA	55.48 ± 2.34	55.17 ± 1.88	58.66 ± 1.86	53.03 ± 1.83
Logistic Regression	57.58 ± 1.34	56.29 ± 1.74	58.70 ± 2.00	53.30 ± 2.24
Gaussian Naive Bayes	53.72 ± 1.88	55.16 ± 2.03	58.65 ± 2.26	51.06 ± 1.44
Kernel Naive Bayes	53.64 ± 2.71	54.73 ± 1.99	57.49 ± 2.20	51.18 ± 1.30
Linear SVM	**60.03 ± 1.72**	56.45 ± 2.33	**59.37 ± 2.05**	**54.70 ± 2.80**
Quadratic SVM	57.98 ± 1.86	55.48 ± 1.67	56.26 ± 2.09	54.04 ± 2.19
Cubic SVM	55.10 ± 1.83	54.03 ± 0.97	54.82 ± 2.09	52.63 ± 1.97
Fine Gaussian SVM	53.82 ± 1.79	52.93 ± 1.68	52.35 ± 1.81	52.49 ± 2.72
Medium Gaussian SVM	59.00 ± 2.57	**56.73 ± 1.80**	58.89 ± 1.85	53.42 ± 2.57
Coarse Gaussian SVM	58.20 ± 1.97	56.48 ± 2.26	59.30 ± 2.06	51.96 ± 2.17

recorded at 56.73 ± 1.80% using Medium Gaussian SVM. In contrast, the third set (CondEn + PCA) enhanced accuracy for specific classifiers, with the Linear SVM being paramount, achieving 59.37±2.05% at its highest. Conversely, employing the fourth set (PLV + clustering coefficient) precipitated a universal decline in overall accuracy across all classifiers, pinpointing 54.70 ± 2.80% for Linear SVM.

V. Discussion and Conclusion

This pilot study introduced a novel BCI baseline that combines LLM-generated labels, particularly from Generative Pre-trained Transformers (GPT-3.5 and GPT-4), with an EEG-based approach for brain state classification and eye-gaze analysis. This is one of the first efforts to use GPT capability for this specialized intersection of cognitive neuroscience and artificial intelligence.

A. Insights From Eye Gaze During Reading

Eye gaze serves as a significant biomarker, holding essential information for understanding the cognitive processes of individuals engaged in task-specific reading activities [47]. In this study, we conducted average fixation analyses on three levels: per individual subject, in relation to specific semantic associations, and at the individual word level. These analyses, leveraging data from 12 participants across eight semantic relations, demonstrate that participants consistently allocate more time to words with high semantic relevance (i.e., keywords) during inference tasks, as corroborated by Appendices A and B.

We also scrutinized single-word fixation statistics across 12 subjects and eight semantic categories within the HRW and LRW groups. Notably, due to missing data for eight relationship instances — with Subject ZGW omitting “JOB, “ZKB missing “WIFE,” and ZPH lacking both “POLITICAL AFFILIATION” and “WIFE” — we included these gaps in the supplementary materials. Our analyses reveal that HRW elicited significantly higher fixation counts compared to LRW as well, shedding light on participants’ comprehension approaches within the corpus.

Table II’s eye gaze metrics distinctly show variable engagement with words based on their semantic relevance. The elevated fixation counts and prolonged gaze durations for HRW reinforce the focus on semantically critical terms. These terms not only captured initial attention, as reflected in the first fixation duration, but also maintained it, evident in the total reading time. Additionally, the shorter single fixation duration on HRW suggests efficient cognitive processing of these terms, while a slight increase in go-past time indicates an extra layer of cognitive effort.

B. Fixation-Related EEG Analysis and Classification

Unlike traditional BCIs, which relied on precise stimulus presentation as timing markers to extract event-related EEG activities such as P300 and Steady-State Visual Evoke Potentials in well-controlled laboratory environments, our approach leveraged fixation onsets to capture EEG signals related to words during natural reading. This implementation significantly enhances the practicality of BCIs for real-world applications.

Fig. 3 presents EEG data related to natural reading, revealing subtle yet discernible differences in brain activity in response to words of varying semantic relevance. The ERP and PSD data across the Pz and Oz channels suggest that HRW may elicit slightly different fixation-related potentials compared to LRW, as indicated by the shaded areas of the graphs. The topographic maps further demonstrate average band power across five frequency bands, with the bottom row highlighting modest differences in power in the occipital region, suggesting a potential disparity in visual processing.

We evaluated the performance of four distinct LLMs to generate robust labels for improving classification outcomes. Our hybrid architecture, combining GPT-3.5 and GPT-4 as word labelers with eye tracking and BCI components, demonstrated robust performance, achieving an accuracy rate exceeding 60% in the classification of word relevance. This enhancement was realized by applying SVMs to three domain-specific features: BP, CondEn combined with PCA, and PLV-based graph theory techniques. Each feature was carefully chosen for its well-established utility in BCI research and its capacity to enhance the SNR. Additionally, we explored the pair-wise coherence of 5-frequency bands but ultimately decided against its use because of its computational complexity, particularly when considering the 105 EEG channels we employed.

The most relevant work to our study is our preliminary experiment detailed in [21], which used seven naive NLP models to determine words ‘similar’ to inference keywords and executed classification using a deep network. However, this approach encountered significant overfitting after 100-150 epochs. The CNN's test accuracy was only marginally better than the LDA model, with the highest test accuracy at 59.3% for cross-subject conditions and 63.3% for within-subject conditions. In contrast, our current work compares 11 non-deep learning methods, using 5-fold validation, both enhancing the robustness of our findings and establishing a new baseline for classifying brain states based on word importance, especially given the high complexity of word-level EEG classification during natural reading comprehension.

Hollenstein et al. [48] used the same ZuCo dataset for EEG cross-subject classification to differentiate between two reading paradigms: normal reading and task-specific reading. However, they applied sentence-level labels for predictions, which diverges from our objective. Duan et al. [49] and Wang and Ji [50] focused on brain-to-text tasks, encoding EEG signals to match word embeddings using language models. Our study aims to discern distinct brain states indicated by EEG biomarkers, whereas theirs primarily translates EEG into words with moderate success.

C. Challenges and Future Work

Despite these advances, the study has several limitations. This study faces challenges because of the "black box" nature of LLMs, particularly in the context of the non-deterministic relation, such as "AWARD," where certain outputted words appear incongruous. This limitation might affect our findings' generalizability and underscore the need for a quantitative assessment to ensure the accuracy and validity of keyword identification.

Additionally, contextual complexities often influence semantic classifications. For example, "gold" acquire distinct semantic relevance when juxtaposed with terms like "medal." The sentences incorporating specific target terms, such as "NATIONALITY" or "WIFE," exhibit a significant disparity in the distribution between HRW and LRW, making them more deterministic. These discrepancies add complexity to the classification of EEG data and introduce the possibility of contamination within the dataset, especially when the meaning of words is most effectively comprehended within the context of phrases rather than in isolation.

This study underscores the potential for more expansive research on elucidating reading-related cognitive behaviors. The promise of integrating LLMs into BCIs also points towards future advancements in reading assistance technologies. While acknowledging its limitations and complexities, our work is an early yet significant contribution, paving the way for more integrated studies to foster a deeper understanding of the multifaceted interplay between neuroscience and computational linguistics.

References

[1] H. C. Wang et al., "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, Aug. 2023.

[2] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[3] M. Abdullah, A. Madain, and Y. Jararweh, "ChatGPT: Fundamentals, applications and social impacts," in *Proc. 9th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Nov. 2022, pp. 1–8.

[4] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.

[5] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI-explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, 2019, Art. no. eaay7120.

[6] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension," *Psychol. Rev.*, vol. 87, no. 4, pp. 329–354, 1980.

[7] K. Rayner, "Eye movements in reading and information processing: 20 years of research.," *Psychol. Bull.*, vol. 124, no. 3, pp. 372–422, 1998.

[8] W. Kintsch, *Comprehension: A Paradigm for Cognition*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[9] M. Binz and E. Schulz, "Using cognitive psychology to understand GPT-3," *Proc. Nat. Acad. Sci. USA*, vol. 120, no. 6, Feb. 2023, Art. no. e2218523120.

[10] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[11] C. Pandarinath et al., "High performance communication by people with paralysis using an intracortical brain–computer interface," *eLife*, vol. 6, Feb. 2017, Art. no. e18554.

[12] D. Shawky and A. Badawi, "Towards a personalized learning experience using reinforcement learning," in *Machine Learning Paradigms: Theory and Application*, 2019, pp. 169–187.

[13] S. Ge, Z. Zhu, B. Wu, and E. S. McConnell, "Technology-based cognitive training and rehabilitation interventions for individuals with mild cognitive impairment: A systematic review," *BMC Geriatrics*, vol. 18, no. 1, pp. 1–19, Dec. 2018.

[14] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP)," *Annu. Rev. Psychol.*, vol. 62, no. 1, pp. 621–647, Jan. 2011.

[15] M. Kutas and S. A. Hillyard, "Reading senseless sentences: Brain potentials reflect semantic incongruity," *Science*, vol. 207, no. 4427, pp. 203–205, Jan. 1980.

[16] W. Kintsch, "The role of knowledge in discourse comprehension: A construction-integration model," *Psychol. Rev.*, vol. 95, no. 2, p. 163, 1988.

[17] J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 255–278, Jan. 2008.

[18] P. N. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA, USA: Harvard Univ. Press, 1983.

[19] D. S. McNamara and J. Magliano, "Toward a comprehensive model of comprehension," in *Psychology of Learning and Motivation*, vol. 51. Academic, 2009, ch. 9, pp. 297–384, doi: 10.1016/S0079-7421(09)51009-2.

[20] X. Liu and Z. Cao, "Enhance reading comprehension from EEG-based brain–computer interface," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 14471, T. Liu, G. Webb, L. Yue, and D. Wang, Eds. Singapore: Springer, 2024, doi: 10.1007/978-981-99-8388-9_44.

[21] Q. Li, *Reading Comprehension Analysis and Prediction Based on EEG and Eye-Tracking Techniques*. San Diego, CA, USA: Univ. of California, San Diego, 2021.

[22] H. Zeng, C. Yang, G. Dai, F. Qin, J. Zhang, and W. Kong, "EEG classification of driver mental states by deep learning," *Cognit. Neurodyn.*, vol. 12, no. 6, pp. 597–606, Dec. 2018.

[23] R. J. Seitz et al., "Valuating other people's emotional face expression: A combined functional magnetic resonance imaging and electroencephalography study," *Neuroscience*, vol. 152, no. 3, pp. 713–722, Mar. 2008.

[24] M. Tanaka, A. Ishii, and Y. Watanabe, "Neural effects of mental fatigue caused by continuous attention load: A magnetoencephalography study," *Brain Res.*, vol. 1561, pp. 60–66, May 2014.

[25] N. Y. AbdulSabur et al., "Neural correlates and network connectivity underlying narrative production and comprehension: A combined fMRI and PET study," *Cortex*, vol. 57, pp. 107–127, Aug. 2014.

[26] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decis. Support Syst.*, vol. 62, pp. 1–10, Jun. 2014.

[27] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading," *Sci. Data*, vol. 5, no. 1, pp. 1–13, Dec. 2018.

[28] H. Brouwer, H. Fitz, and J. Hoeks, "Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension," *Brain Res.*, vol. 1446, pp. 127–143, Mar. 2012.

[29] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.

[30] A. Culotta, A. McCallum, and J. Betz, "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," in *Proc. main Conf. Human Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2006, pp. 296–303.

[31] T.-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Analyzing and visualizing single-trial event-related potentials," in *Advances in Neural Information Processing Systems*, vol. 11, M. Kearns, S. Solla, and D. Cohn, Eds., Cambridge, MA, USA: MIT Press, 1998.

[32] I. Jayarathne, M. Cohen, and S. Amarakeerthi, "Person identification from EEG using various machine learning techniques with inter-hemispheric amplitude ratio," *PLOS one*, vol. 15, no. 9, 2020, Art. no. e0238872, doi: 10.1371/journal.pone.0238872.

[33] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010.

[34] Y. Zhang, Y. Liao, Y. Zhang, and L. Huang, "Emergency braking intention detect system based on K-order propagation number algorithm: A network perspective," *Brain Sci.*, vol. 11, no. 11, p. 1424, Oct. 2021.

[35] W. Ding, Y. Zhang, and L. Huang, "Using a novel functional brain network approach to locate important nodes for working memory tasks," *Int. J. Environ. Res. Public Health*, vol. 19, no. 6, p. 3564, Mar. 2022.

[36] Y. Chen, Y. Zhang, W. Ding, F. Cui, and L. Huang, "Research on working memory states based on weighted k-order propagation number algorithm: An EEG perspective," *J. Sensors*, vol. 2022, pp. 1–10, Jul. 2022.

[37] S. Aydore, D. Pantazis, and R. M. Leahy, "A note on the phase locking value and its properties," *NeuroImage*, vol. 74, pp. 231–244, Jul. 2013.

[38] A. Fornito, A. Zalesky, and E. Bullmore, *Fundamentals of Brain Network Analysis*. New York, NY, USA: Academic, 2016.

[39] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nature Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, Mar. 2009.

[40] K.-J. Chiang, S. Dong, C.-K. Cheng, and T.-P. Jung, "Using EEG signals to assess workload during memory retrieval in a real-world scenario," *J. Neural Eng.*, vol. 20, no. 3, Jun. 2023, Art. no. 036010.

[41] F. Lotte et al., "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Apr. 2018, Art. no. 031005, doi: 10.1088/1741-2552/aab2f2.

[42] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 056013.

[43] Y. Li, R. Yin, H. Park, Y. Kim, and P. Panda, "Wearable-based human activity recognition with spatio-temporal spiking neural networks," 2022, *arXiv:2212.02233*.

[44] V. Wyart and C. Tallon-Baudry, "How ongoing fluctuations in human visual cortex predict perceptual awareness: Baseline shift versus decision bias," *J. Neurosci.*, vol. 29, no. 27, pp. 8715–8725, Jul. 2009.

[45] C. Tallon-Baudry, O. Bertrand, M.-A. Hénaff, J. Isnard, and C. Fischer, "Attention modulates gamma-band oscillations differently in the human lateral occipital cortex and fusiform gyrus," *Cerebral Cortex*, vol. 15, no. 5, pp. 654–662, May 2005, doi: 10.1093/cercor/bhh167.

[46] S. Palva, S. Kulashekhar, M. Hämäläinen, and J. M. Palva, "Localization of cortical phase and amplitude dynamics during visual working memory encoding and retention," *J. Neurosci.*, vol. 31, no. 13, pp. 5013–5025, Mar. 2011.

[47] S.-C. Chen, H.-C. She, M.-H. Chuang, J.-Y. Wu, J.-L. Tsai, and T.-P. Jung, "Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities," *Comput. Educ.*, vol. 74, pp. 61–72, May 2014.

[48] N. Hollenstein et al., "The ZuCo benchmark on cross-subject reading task classification with EEG and eye-tracking data," *Frontiers Psychol.*, vol. 13, Jan. 2023, Art. no. 1028824.

[49] Y. Duan, J. Zhou, Z. Wang, Y.-K. Wang, and C.-T. Lin, "DeWave: Discrete EEG waves encoding for brain dynamics to text translation," 2023, *arXiv:2309.14030*.

[50] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 5, pp. 5350–5358.