Preprints.org

Article

Not peer-reviewed version

Exploring the Cognitive Neural Basis of Factuality in Abstractive Text Summarization Models: Interpretable Insights from EEG Signals

Zhejun Zhang , Yingqi Zhu , Yubo Zheng , Yingying Luo , Hengyi Shao , Shaoting Guo , Liang Dong , Lin Zhang * , Lei Li *

Posted Date: 27 November 2023

doi: 10.20944/preprints202311.1717.v1

Keywords: natural language processing (NLP); abstractive summarization (ABS); factual extraction; Electroencephalography (EEG); Representational similarity analysis (RSA)

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Exploring the Cognitive Neural Basis of Factuality in Abstractive Text Summarization Models: Interpretable Insights from EEG Signals

Zhejun Zhang, Yingqi Zhu, Yubo Zheng, Yingying Luo, Hengyi Shao, Shaoting Guo, Liang Dong, Lin Zhang* and Lei Li *

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China; zhejun.zhang@bupt.edu.cn (Z.Z.); 809514302@qq.com (QY.Z.); zyb@bupt.edu.cn (YB.Z.); luoyy@bupt.edu.cn (Y.L.); jsjjkfq740118@bupt.edu.cn (H.S.); guost@bupt.edu.cn (S.G.); dongliang@bupt.edu.cn (L.D.)

* Correspondence: zhanglin@bupt.edu.cn (L.Z.); leili@bupt.edu.cn (L.L.)

Abstract: (1) Background: Information overload challenges decision-making in the Industry 4.0 era. While Natural Language Processing (NLP), especially Automatic Text Summarization (ATS), offers solutions, issues with factual accuracy persist. This research bridges cognitive neuroscience and NLP, aiming to improve model interpretability. (2) Methods: This research examined four fact extraction techniques: dependency relation, named entity recognition, part-of-speech tagging, and TF-IDF, in order to explore their correlation with human EEG signals. Representational Similarity Analysis (RSA) was applied to gauge the relationship between language models and brain activity. (3) Results: Named entity recognition showed the highest sensitivity to EEG signals, marking the most significant differentiation between factual and non-factual words with a score of -0.99. The dependency relation followed with -0.90, while part-of-speech tagging and TF-IDF resulted in 0.07 and -0.52, respectively. Deep language models such as GloVe, BERT, and GPT-2 exhibited noticeable influences on RSA scores, highlighting the nuanced interplay between brain activity and these models. (4) Conclusions: Our findings emphasize the crucial role of named entity recognition and dependency relations in fact extraction and demonstrate the independent effects of different models and TOIs on RSA scores. These insights aim to refine algorithms to reflect human text processing better, thereby enhancing ATS models' factual integrity.

Keywords: natural language processing (NLP); abstractive summarization (ABS); factual extraction; electroencephalography (EEG); representational similarity analysis (RSA)

1. Introduction

In the era of Industry 4.0, information technology rapidly promotes industrial transformation but simultaneously leads to information overload, exposing people to vast amounts of textual information. The challenge of information overload intensifies when there is a need to quickly extract critical information from tedious text for decision-making, highlighting data quality and accuracy. Natural Language Processing (NLP), a fundamental branch of Artificial Intelligence (AI), offers a viable solution to the problem. Specifically, Automatic Text Summarization (ATS), an NLP technology, has been widely applied in news extraction, academic research, business reports, legal analysis, content recommendation, and various other fields.[1,2] In recent years, transformer architecture has gained significant popularity, leading to notable advancements in fluency and readability of Abstractive Summarization (ABS) for language models such as GPT, BERT, and BART. However, certain limitations remain regarding the factual accuracy of summaries.[3] The chatbot ChatGPT cautions users that "ChatGPT may produce inaccurate information about people, places, or facts" beneath its conversation interface. Factual errors include misrepresenting or omitting information, distortion of logical relationships, and erroneous interpretation of facts.[4] In domains with high demands for information accuracy, like healthcare, law, and finance, factual errors could lead to severe consequences.

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

In many NLP tasks, neural networks are the most advanced machine learning methods. However, their interpretability is frequently challenged.[5] Applying word vectors and lacking structured textual data inputs increase model opacity, rendering their outputs challenging to interpret.[6] Enhancing model interpretability is crucial as it facilitates comprehension of the functioning of "black box" models, thereby enabling their improvement.[7] ABS entails Natural Language Understanding (NLU) and Natural Language Generation (NLG), making it complex. Improved interpretability aids researchers in comprehending the process of summarization generation, encompassing aspects such as fact retrieval within the model, determination of dependency relations, and identification of potential erroneous entity replacements. The enhanced interpretability enables a targeted reduction in factual errors while preserving the integrity of the original text information to enhance the summary compression rate, thereby improving the summaries' accuracy, fluency, and readability.[8]

In recent years, the intersection of computational modeling and cognitive neuroscience has garnered increasing attention. Relevant research contributes to understanding how language and information are processed in the human brain, offering novel perspectives and methodologies for enhancing and optimizing existing NLP models.[9,10] For instance, by emulating cognitive characteristics observed in human linguistics tasks, significant advancements have been achieved in named entity recognition and other related NLP tasks.[7] However, it is worth noting that no prior studies have explored text summarization from a cognitive neuroscience perspective. Given the significance of text summarization in facilitating quick information extraction and efficient knowledge management, as well as the current limitations regarding factual accuracy within existing models, this study aims to apply theories and methods from cognitive neuroscience to establish correlations between human brain activities and ABS model performance. This approach will enhance interpretability in text summarization while advancing interdisciplinary research. Specifically, the main contributions of this study are:

1. This study pioneers using EEG signals from a cognitive neuroscience perspective to investigate factual issues of ABS, offering novel insights into the relationship between models and human brain activity in language tasks.

2. This study compares variations in EEG signals corresponding to factual word phrases and non-factual word groups obtained through different fact term extraction methods (dependency relation, named entity, part-of-speech tagging, and TF-IDF). It was found that the distinctions in EEG were most significant for factual and non-factual word groups using named entity recognition and dependency relations. This lays a foundation for integrating extraction methods to better simulate human text processing.

3. This study employs Representational Similarity Analysis (RSA) to compare the correlation between typical deep language models (GloVe, BERT, GPT-2) and human brain activities, revealing significant differences in RSA scores among different models and periods of human brain activity under specific conditions. These findings suggest potential adjustments to enhance the model's resemblance to the functioning of the human brain and facilitate a deeper understanding of mechanisms involved in language processing tasks within the human brain.

2. Related Works

2.1. Abstractive Text Summarization

The advancement of automatic text summarization technology facilitates the generation of concise and coherent summaries that effectively address information overload.[11,12] Since 2020, Transformer models and their variants, such as GPT, BERT, and BART, have demonstrated remarkable performance in NLP, particularly in generative tasks.[3] Unlike extractive summarization, which directly extracts information from the source text, abstractive summarization requires a deeper understanding of textual semantics and employs NLG algorithms to rephrase critical points.[13] This approach makes abstractive summarization more similar to human-generated summarization and significantly improves fluency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

and readability.[14] The complicated generation process, however, presents a prominent challenge: factual accuracy. During summary generation, errors in entity replacement or inaccuracies in logical relationships are expected. Such factual errors are challenging to detect or rectify using current statistical metrics.[15]

Defining "what is a fact" is a crucial foundation for addressing the issue of factual accuracy, yet no unified definition for it currently exists. Before 2020, most studies regarded relational triples as the fundamental form of facts, using them to improve the factual awareness of language models and precision. However, comprehensive factual triples are not always extractable. Existing research often involves the introduction of numerous auxiliary virtual entities and additional triples, conversions that contribute to the complexity of predicting links for two or more "arcs".[16] In recent years, broader dimensions of fact definitions have been proposed. Some researchers defined facts based on the keywords in the original text, such as the TF-IDF method.[17] This approach is essentially statistical, assessing the significance of a word in a document or corpus, enabling an intuitive identification of the text's theme. However, such statistical methods fail to comprehend the relations between identified words and others. Part-of-speech tagging has also been utilized for fact extraction, assigning specific functional labels to each word in a text, such as nouns, verbs, or adjectives.[18] Different parts of speech may represent various factual elements and tagging them helps identify their specific roles in sentences, which applies to texts of diverse types and styles. Nonetheless, this method might be overly simplistic, potentially overlooking more intricate semantic relationships. Entity recognition[15] has also been employed for fact extraction, discerning items with specific meanings from the text, and viewing entities as facts, proving valid in downstream tasks like text summarization and information retrieval with minimal cumulative error. However, entity recognition might neglect important information unrelated to entities. Some researchers employ dependency relations as a fact extraction method to uncover intricate relations between text lexemes. Dependency relations aid in understanding the primary structure of texts, transforming input sentences into labeled tuples, and extracting tuples associated with predicate lexemes.[19] However, these relations mainly capture inter-word relationships. In complex sentences, if the dependency path between two words is too lengthy or contains numerous nodes, the facts derived from these paths might fail to fully represent the sentence's intended meaning.

Understanding the characteristics of human linguistic cognition is of paramount importance. Some researchers have significantly enhanced the performance of abstractive summarization models by simulating features of human linguistic cognition.[20] When humans read and comprehend texts, they typically rely on various cues to discern which information is pivotal or significant. These cues often stem from individuals' expectations and knowledge background, bearing a certain degree of subjectivity, and might operate subconsciously, making them challenging to emulate directly through computational models. EEG signals during natural reading might encompass elusive information, such as sentence structures or specific named entities. Appling EEG signals from natural reading to selecting or designing fact extraction methods that align with human cognitive characteristics or guiding extraction methods' amalgamation hold significant implications for producing accurate and reliable text summaries.

2.2. Linguistic Cognition and Neural Network Modeling

ABS encompasses NLP and NLG. Therefore, understanding the cognitive mechanisms of language comprehension and generation in the human brain may offer valuable insights into the ABS research. Neural signals from the human brain bridge the gap between the information processes mechanisms of cognition and models. EEG measures brain electrical activity by placing electrodes on the scalp, enabling the observation and analysis of real-time brain activity during specific tasks. With the capability to capture millisecond-level changes in brain activity, EEG boasts a high temporal resolution, making it apt for studying rapid cognitive processes such as language comprehension and generation. Although its spatial resolution is lower than functional Magnetic Resonance Imaging (fMRI), EEG can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

still monitor activity at multiple scalp locations and identify electrode positions related to specific tasks. As the brain comprehends text when the meaning of a word mismatches the overall sentence meaning, a neural potential peak often emerges, characterized by a negative voltage shift between 300ms and 500ms. This Event-Related Potential (ERP) is termed the N400 effect.[21] The N400 predominantly resides in the central and posterior regions of the brain, especially in the middle and rear sections of the temporal lobe. In contrast to semantics, when dealing with the processing of intricate syntactic structures, the brain manifests the P600 effect, a positive voltage shift that peaks approximately 600ms post-stimulus.[22] Spatially, it is associated with the brain's left temporal and parietal regions. The N400 and P600 effects are pivotal tools in cognitive neuroscience for studying language processing, delineating EEG response patterns within specific time windows related to semantic and syntactic processing, respectively.

In recent years, the interdisciplinary research between computational models and cognitive neuroscience in NLP has garnered increasing attention. Researchers have sought to intertwine these two domains from multiple perspectives. For example, Lamprou et al.[10] and Ikhwantri et al.[9] have elucidated the language processing of neural networks from a neurolinguistic viewpoint and directed the training of neural networks based on the human brain's text-processing mechanisms. The objectives of these studies encompass understanding the human brain's operational mechanisms and optimizing the performance of NLP models. In terms of model optimization inspired by cognitive neuroscience, Y. Chen et al.[11] introduced a controlled attention mechanism for named entity recognition, which exhibited exemplary performance across multiple datasets. Besides, Ren et al.[23] successfully integrated cognitive signals into neural network NLP models through the CogAlign method, with experimental results indicating its efficacy in enhancing model performance. To gain a deeper comprehension of how to map models onto human brain activity, Oseki & Asahara[24] designed a method to obtain EEG signals from participants during natural reading of a specific corpus and used the processed EEG signals to annotate various levels of the corpus. Oota et al.[25] further discerned that representations learned from different NLP tasks respectively interpret the brain responses to speech reading and listening: representations from semantic tasks (such as paraphrase generation, text summarization, and natural language inference) are more pertinent for listening comprehension, while those from syntactic tasks (like coreference resolution and shallow syntactic parsing) are more pertinent for reading comprehension. Additionally, some researchers have ventured from cross-modal and multilingual perspectives. For instance, leveraging cross-modal transfer learning, Antonello et al.[26] discovered a low-dimensional structure that seamlessly bridges various linguistic representations learned by different language models, including word embeddings and tasks related to syntactic and semantic processing. This low-dimensional representation embedding reflects the hierarchical structure of language processing and can predict fMRI responses elicited by linguistic stimuli. On the other hand, Giorgi et al.[27], approaching from a developmental neuroscience perspective, introduced a neural network architecture designed to learn multiple languages concurrently.

RSA is a prevalent technique to evaluate the relationship between deep language models and neural activity.[28] Lenci et al.[29] highlighted that RSA is particularly suited for datasets that are challenging to compare directly, such as neural brain activity and internal representations of machine learning models. The core concept of RSA involves transforming raw data or model representations into a common similarity space, typically achieved by computing similarity matrices. The application of RSA unveils which sections of the neural network model are most similar to brain neural signals. Moreover, it can also facilitate cross-validation between brain data and multiple computational models to determine which model best accounts for brain activity.

In summary, integrating cognitive neuroscience and computational models allows for a deeper understanding of the linguistic processing mechanisms within the human brain. It facilitates the optimization and enhancement of existing NLP models. Research in this interdisciplinary field paves new avenues for future endeavors in NLP and neuroscience. However, despite various NLP tasks covered in this cross-disciplinary research, such as named entity recognition and natural

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

language inference, there remains an absence of studies specifically applying insights from cognitive neuroscience to elucidate or optimize text summarization tasks.

3. Materials and Methods

3.1. Participants

This study recruited 14 valid participants, nine males and five females, with an average age of 22.64±2.90 years. All participants met the following criteria: (a) they had passed the English CET6 examination; (b) they had not dyed their hair in the past two months; (c) they had normal or corrected to normal vision, with no visual impairments such as color blindness or color weakness; (d) they had no history of psychiatric illnesses or mental disorders and no language or motor impairments; (e) they had not experienced physical discomfort (e.g., cold) in the past week and ensured adequate rest the day before the experiment.

3.2. Apparatus and Materials

This study used the BEATS system to collect EEG data.[30] This system comprises an analog front end, a microprocessor, and a software platform that offers an integrated solution from hardware to software. It is capable of wirelessly collecting 32-channel EEG signals.

The experimental materials for this study were selected from the Factuality Evaluation Benchmark (FRANK) dataset.[4] This dataset serves as an abstract summarization benchmark for factual evaluation metrics. It encompasses 2250 summaries generated by nine models on two datasets (CNN/DM and XSum) and manual annotations. Each sentence within a summary is annotated for the presence of factual errors and the type of those errors. These types are determined based on a factual error classification system, which in turn is derived from semantic frames and theories of semantic coherence. The criteria for material selection in this study included: (a) annotations in the model output indicated factual errors, highlighting the necessity of model improvement; (b) To avoid fatigue effects, each chosen article did not exceed 200 words. Furthermore, no more than seven articles were used in each experiment, ensuring that the experimental duration remained below 1 hour; (c) the infrequent vocabulary and proper nouns were minimal.

The details of the selected materials are presented in Table 1. A total of 7 English short texts were chosen for the study. After merging all texts and tokenizing using the Python spaCy, 1220 tokens were obtained.

Table 1. Selection of experimental materials.

Article hash	Count of sentences	Count of words	Count of tokens
32143053	10	154	180
38329319	11	200	242
32457391	7	152	176
33652722	6	135	148
31566848	6	128	143
31920236	7	139	160
38595401	6	138	171

The experimental procedure was programmed using Psychopy (version 2022.2.5). Psychopy is an open source psychological experiment software suitable for crafting various experimental tasks.[31,32] For this study, a continuous text reading task was devised to record both stimulus presentation times and participant behavioral data. In subsequent sections, unless expressly stated otherwise, any mention of the term "word" refers to a "token."

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

3.3. Experimental Procedure

The experimental procedure could be divided into three stages: preparation, pre-experiment, and formal experiment. The study was conducted in a sealed laboratory, where the lights were turned off, and external noise was isolated during the formal experiment. In the preparation stage, the experimenter explained the experiment's purpose, process, and potential risks to the participants, ensuring that all participants read and voluntarily signed an informed consent form. Subsequently, the experimenter cleaned participants' scalps to ensure optimal electrode adhesion. Then, the experimenter fitted the EEG cap onto the participant's head and applied electrode gel to enhance the contact quality between the electrodes and the scalp. Lastly, the functionality of the EEG equipment and recording system was checked to ensure the correct setup of system parameters and that all electrode channels were connected correctly, thus ensuring the acquisition of the required data.

In the pre-experiment stage, participants followed a procedure similar to the formal experiment but only needed to read a shorter text to ensure they understood the experimental task. During the formal experiment, as illustrated in Figure 1, participants sat in front of a computer screen, gazing horizontally at it, and were instructed to minimize head movement throughout the session. Initially, a fixation cross appeared at the center of the screen for 5 seconds, followed by a pre-selected text presented word-by-word, with each word displayed for 1.5 seconds. Punctuation marks were not presented separately. After reading an entire text, participants were required to answer three comprehension questions to assess their understanding of the text. Only results from participants with a final accuracy rate above 50% were considered valid. Following the reading of each text, participants were given a break. If any discomfort arose during the experiment, participants could press the "Esc" key at any time to exit the experimental program. Throughout the experiment, one experimenter guided participants through the tasks while another monitored real-time EEG signals, saving the EEG data corresponding to each participant's reading of each text.

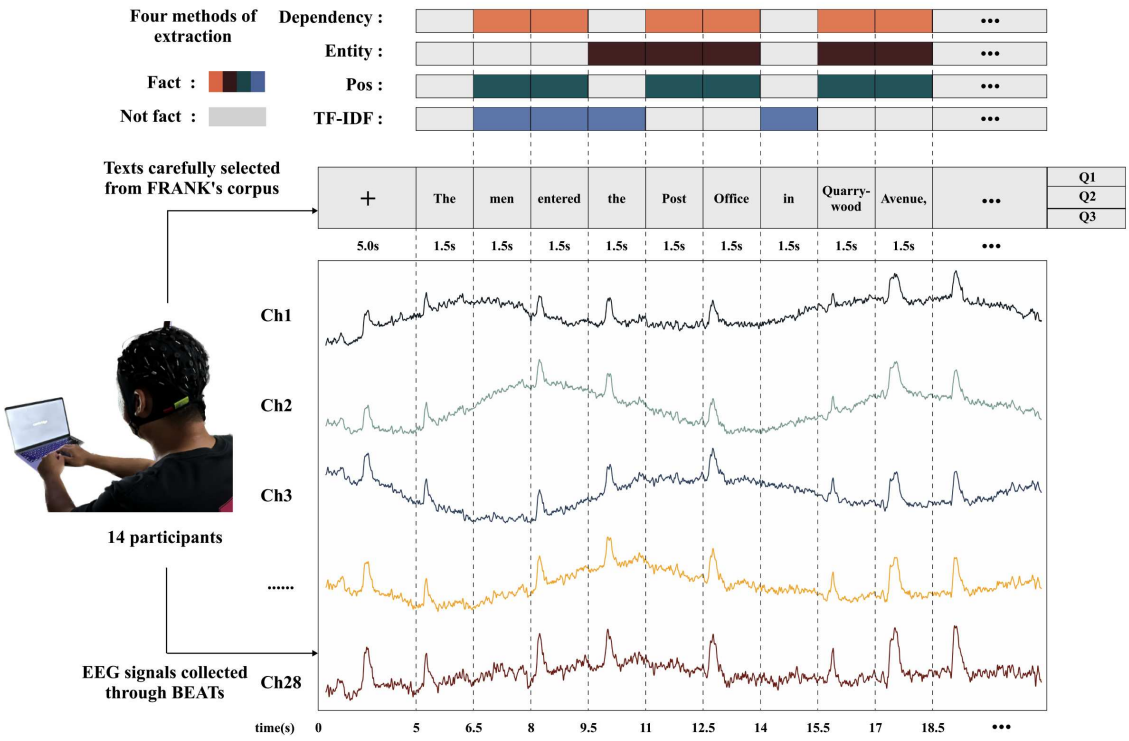

Figure 1. Experimental procedure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

3.4. Collection and Preprocessing of EEG Data

The EEG signals in this study were collected from 28 electrodes embedded in an elastic cap arranged according to the 10/20 system, as illustrated in Figure 2. Throughout the process of collecting, the impedance of each electrode was maintained below 10kΩ, and the sampling rate for the EEG signals was set at 250Hz. Preprocessing was conducted using Python's SciPy. During the preprocessing phase, the low-pass filter of EEG data had a cutoff frequency set at 30Hz, while the high-pass filter's cutoff was set at 0.1Hz. Filtering was executed independently on each channel. Subsequently, a reference transformation was applied to the EEG data to eliminate common noise and background signals across channels. This step involved calculating the average value across all channels at each time point and subtracting this average from each channel's signal. Finally, baseline correction was executed by determining the mean value for every channel over the entire recording duration and subtracting this mean from each respective time point within that channel.

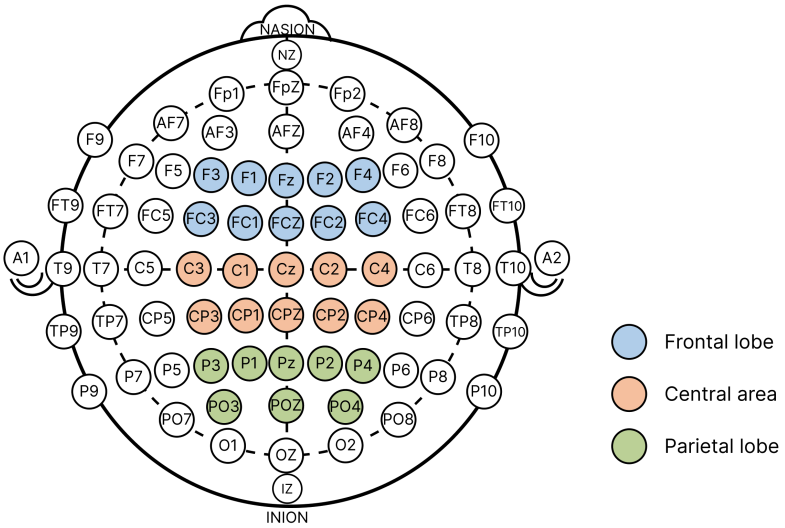

Figure 2. EEG channels selected in the study.

3.5. Metrics Selection and Data Analysis

3.5.1. Sensitivity of EEG Signals to Different Fact-Extraction Methods

In the current study, Python's spaCy library was employed for preprocessing the corpus. As discussed in related works, four extraction methods (dependency, entity, pos, and TF-IDF) were selected. For clarity, any subsequent mention of the term "word" in this context refers to a "token."

For dependency relation extraction (dependency), spaCy initially identified dependency relations among all words. Then it selected labels representing relationships between words, including: 'ROOT,' 'nsubj,' 'nsubjpass,' 'compound,' 'poss,' 'pcomp,' 'ccomp,' 'conj,' 'relcl,' 'dobj,' 'pobj,' 'iobj,' 'appos,' and 'acl,' totaling 14 labels. These labels were considered capable of extracting fact-related information. As for named entity recognition (entity), spaCy tagged proper nouns in the text, such as names of people, places, organizations, and dates. Part-of-speech tagging (pos) was also executed using spaCy, extracting nouns and verbs from the text. When employing the TF-IDF method for extraction, Term Frequency (TF) and Inverse Document Frequency (IDF) were primarily calculated and then multiplied, as presented in the subsequent formula (1-3). The documents selected for IDF calculation included all entries from the FRANK dataset with word counts below 1000. When calculating TF, all words were considered equally important. However, common words like "the" and "a" might have appeared frequently but might not have been crucial. Thus, we needed to downplay the weight of these words. IDF served as a method to reduce the weight of these words.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document} \tag{1}$$

$$IDF(t) = \log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right) \tag{2}$$

$$TF\text{-}IDF(t) = TF(t) \times IDF(t) \tag{3}$$

Cosine similarity can be employed to assess the similarity between vectors. Researchers utilize the cosine similarity of word weight vectors to compare the degree of similarity between documents or the cosine similarity of word vectors to compute the similarity between different words.[33] Inspired by prior studies, this research considers each token's corresponding EEG signal as an EEG vector, using their cosine similarity to assess the sensitivity of EEG signals to different factual word extraction methods. Specifically, after preprocessing, the EEG data from reading the text in this study was saved with a structure of $(n,\ 28,\ 375)$. n was the number of tokens of the corresponding text after tokenization by spaCy. Twenty-eight represented the number of channels for EEG signal collection, and 375 was the number of potentials collected within 1.5 seconds for each word. After averaging across the 28 channels, the data structure becomes $(n,\ 375)$, suggesting each token corresponds to a 375-dimensional EEG vector. In addition to the full 1.5 seconds, this study also established two TOI intervals: the 250-500ms interval, which might contain the N400 ERP component, and the 500-1000ms interval, which might contain the P600 ERP component.

The cosine similarity between the EEG vectors of factual and non-factual word groups is calculated as presented in Equation (4). Here, c_1 and c_2 represent the centroids of the vectors for the factual and non-factual word groups, respectively. The centroids are determined by taking the mean of all vectors within the factual and non-factual word groups. Unlike the similarity between word vectors, which is always non-negative, the similarity range for EEG vectors lies between -1 and 1. A smaller angle between vectors indicates a closer directionality, making the cosine similarity approach 1. Conversely, a larger angle suggests divergent directions, bringing the cosine similarity closer to -1. When the angle is 90 degrees, the cosine similarity is 0, signifying that the vectors are orthogonal and unrelated. The method for calculating the cosine similarity between EEG vectors within either the factual or non-factual word groups is depicted in Equation (5). In this Equation, n is the total number of words within the word group, while v_1 and v_2 are the EEG vectors for the i-th and j-th words, respectively.

$$Inter\text{-}class\ Cosine\ Similarity = \frac{c_1 \cdot c_2}{\|c_1\|_2 \times \|c_2\|_2} \tag{4}$$

$$Intra\text{-}class\ Cosine\ Similarity = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{v_i \cdot v_j}{\|v_i\|_2 \times \|v_j\|_2} \tag{5}$$

3.5.2. Correlation between Human Brain and Models

This study employed RSA to assess the similarity between model and human brain activity. The core concept of RSA is that when representational systems (e.g., neural network models or the human brain) receive many inputs, they measure the activity patterns generated by each input. By calculating the response similarity for each possible pair of inputs, one can construct a representational similarity matrix, which encapsulates the internal structure of the representational system. In this research, with 1220 tokens in the text, the analysis of the representational similarity between the model and human brain activity was divided into three steps, as depicted in Figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

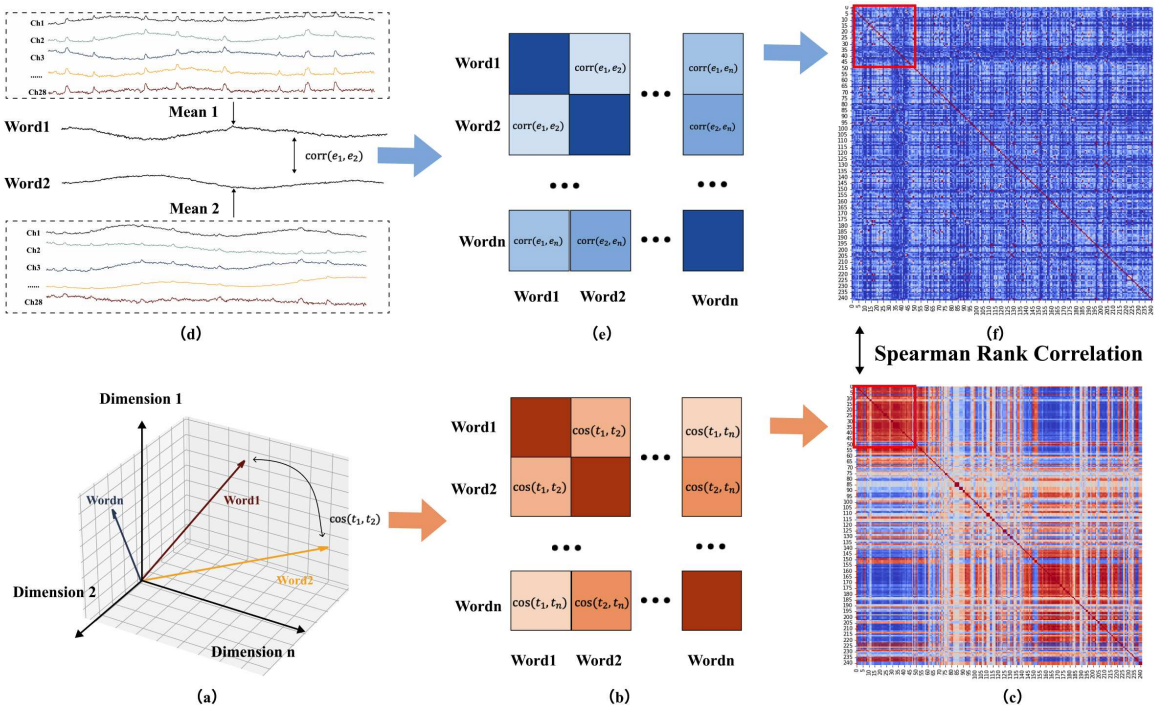

Figure 3. Steps for Calculating Representational Similarity.

The first step was to compute the representational similarity matrix for the deep language model. Initially, the cosine similarity between the word vectors of the 1220 tokens in the article was calculated, as described in Equation (6), where $v(t_i)$ represented the word vector of the i-th token. These similarities were arranged into a matrix, with the calculation process illustrated in Figures 3(a) to 3(c). This similarity matrix's upper and lower triangles were mirror images of each other, with its diagonal representing the similarity of a word to itself.

$$Cosine\ Similarity(t_i, t_j) = \frac{v(t_i) \cdot v(t_j)}{\|v(t_i)\|_2 \times \|v(t_j)\|_2} \tag{6}$$

The second step involves computing the representational similarity matrix for the EEG signals. Similarly, the Pearson correlation coefficients between the EEG vectors corresponding to the 1220 tokens in the article are calculated. In Equation (7), $e(t_{i,k})$ represented the EEG signal of the i-th token at the k-th sampling point, and $\bar{e}(t_i)$ denoted the average EEG signal of the i-th token, averaged across 375 sampling points. These similarities were then arranged into a matrix in the same order as in the first step, with the calculation process illustrated in Figures 3(d) to 3(f).

$$Pearson's\ corr(e_i, e_j) = \frac{\sum_{k=1}^{375}(e(t_{i,k}) - \bar{e}(t_i))(e(t_{j,k}) - \bar{e}(t_j))}{\sqrt{\sum_{k=1}^{375}(e(t_{i,k}) - \bar{e}(t_i))^2}\sqrt{\sum_{k=1}^{375}(e(t_{j,k}) - \bar{e}(t_j))^2}} \tag{7}$$

The third step involved computing the Spearman rank correlation coefficient between the two matrices, as described in Equation (8), where d_i^2 represents the squared difference in ranks for each element between the two matrices, and n is the number of elements in the matrix. This correlation coefficient reflects the representational similarity between the model and human brain activities.[34]

$$Spearman\ Rank\ Correlation = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{8}$$

In this study, we selected "GloVE,"[35] "BERT,"[36] and "GPT-2"[37] models to compare with human brain activity related to language comprehension. "GloVE" is a word embedding method based on global word frequency statistics and is considered a relatively traditional model. Both "BERT" and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

"GPT-2" are models based on the Transformer architecture, and they excel in handling long-distance dependencies and capturing complex structures within sentences. However, "BERT" is a bidirectional model primarily designed for understanding context, while "GPT-2" is a generative model often used for text generation. By comparing these three distinct types of models, we aimed to gain a more comprehensive understanding of the similarities and differences between models and the human brain in processing language.

4. Results

4.1. Sensitivity of EEG Signals to Different Fact-Extraction Methods

4.1.1. Based on cosine similarity

This section reports the inter-class similarities of the EEG vectors corresponding to the fact word groups and non-fact word groups obtained from four fact word extraction methods and the intra-class similarities for each group. The quantities of fact and non-fact words extracted using the four methods (dependency, entity, pos, and TF-IDF) are shown in Table 2.

Table 2. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Extraction Methods	Number of Fact Words	Number of Non-Fact Words
Dependency	522	698
Entity	250	970
Pos	563	657
TF-IDF	478	742

Based on the description in the "Methods" section, each 1220 word was recorded with 1.5 seconds of EEG signal. After averaging across the 28 channels, each word corresponds to a 375-dimensional EEG vector. When computing the cosine similarity of EEG vectors for fact word groups and non-fact word groups obtained from different extraction methods, three distinct time segments of the 1.5-second EEG signal were analyzed for each ex-traction method. These three segments are as follows: (1) the entire 1.5-second duration, encompassing 375 sampling points, denoted as "Overall"; (2) the time window potentially capturing the N400 ERP component, ranging from 250ms to 500ms, which includes 62 sampling points, denoted as "N400"; and (3) the time window potentially capturing the P600 ERP component, ranging from 500ms to 1000ms, containing 125 sampling points, denoted as "P600".

Firstly, the inter-class cosine similarity between the EEG vectors of fact word groups and non-fact word groups obtained from the four extraction methods was computed, with results in Table 3.

Table 3. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

TOI	Dependency	Entity	Pos	TF-IDF
Overall	-0.90	-0.99	0.07	-0.52
N400	-0.94	-1.0	0.88	-0.45
P600	-0.94	-1.0	0.89	0.44

Overall: 0-1500ms; N400: 250-500ms; P600: 500-1000ms

It can be observed that the inter-class cosine similarity between the EEG vectors of fact word groups and non-fact word groups extracted using the Entity and Dependency methods is close to -1. The result indicates that the EEG signals for these fact word groups and non-fact word groups, as identified by these two methods, are highly dissimilar in direction and exhibit clear differentiation. In contrast, the EEG signals' differentiation between the two groups of words extracted based on pos

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

is not pronounced, with the vectors even displaying a degree of positive correlation. Furthermore, the two groups of words extracted using TF-IDF exhibited significant differences in the N400 (often indicative of semantic comprehension) and P600 (typically representing syntactic processing) time windows, suggesting that the TF-IDF method captures some semantic information (as evidenced by the negative similarity in the N400 time window) but overlooks crucial syntactic information (evidenced by the positive similarity in the P600 window).

Subsequently, the intra-group similarity for factual word groups obtained from the four extraction methods was computed, with the results in Table 4.

Table 4. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

TOI	Dependency	Entity	Pos	TF-IDF
Overall	0.0023	-0.0007	0.0018	0.0150
N400	0.0038	-0.0022	0.0024	0.0175
P600	0.0023	-0.0060	0.0021	0.0144

Overall: 0-1500ms; N400: 250-500ms; P600: 500-1000ms

All extraction methods yielded a very low intra-group cosine similarity for the factual word group, nearing 0. The result indicates a minimal correlation between the EEG vectors corresponding to the factual words extracted by these four methods. The angles between the vectors are nearly orthogonal, reflecting the brain's highly independent interpretation of each word within the factual word group.

Lastly, the intra-group similarity for non-factual word groups obtained from the four extraction methods was computed, with the results presented in Table 5.

Table 5. This is a table caption. Tables should be placed in the main text near to the first time they are cited

TOI	Dependency	Entity	Pos	TF-IDF
Overall	0.0094	0.0140	0.0106	0.0021
N400	0.0087	0.0140	0.0109	0.0020
P600	0.0107	0.0156	0.0120	0.0031

Overall: 0-1500ms; N400: 250-500ms; P600: 500-1000ms

All extraction methods resulted in a very low intra-group cosine similarity for the non-factual word group, again approaching 0. The result indicates a minimal correlation between the EEG vectors corresponding to the non-factual words extracted by these four methods, and the brain's comprehension of each word within the non-factual word set is highly independent.

4.1.2. Based on EEG signals

In order to visually illustrate the disparities in brain activity between factual and non-factual words acquired through four extraction methods (dependency, entity, pos, TF-IDF), corresponding EEG signal curves were graphed, as depicted in Figures 4– 7 . Figure 4 corresponds to the four extraction techniques. The curves represent the average EEG signals over 1.5 seconds for 14 participants reading all factual or non-factual words. The semi-transparent regions on either side of the curves indicate the voltage's standard error (SE) at corresponding time points. In the legend, "N" represents the number of EEG signals used for computing the average. For instance, in Figure 4, 7308 indicates that 14 participants read 522 factual words. For each extraction method, four subfigures display the voltage from different channels (Area of Interest, AOI). From left to right, these are all 28 channels (Overall), ten channels located above the frontal lobe (Frontal), ten channels above the central region (Central), and eight channels above the parietal lobe (Parietal).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

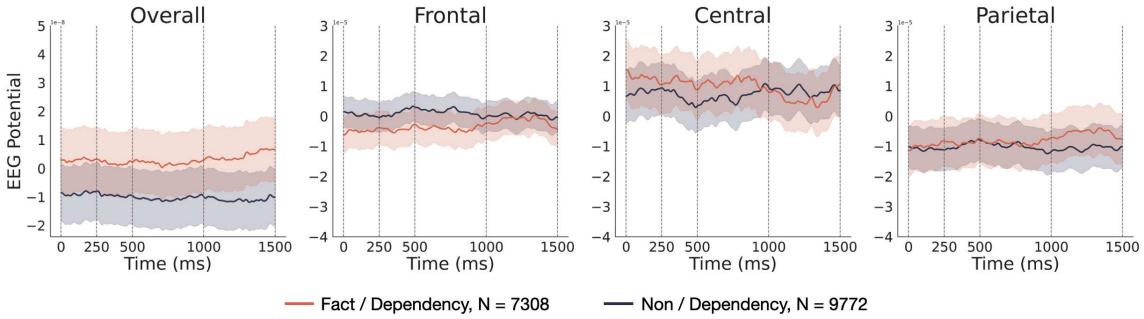

Figure 4. EEG signals obtained using the extraction method of dependency relations.

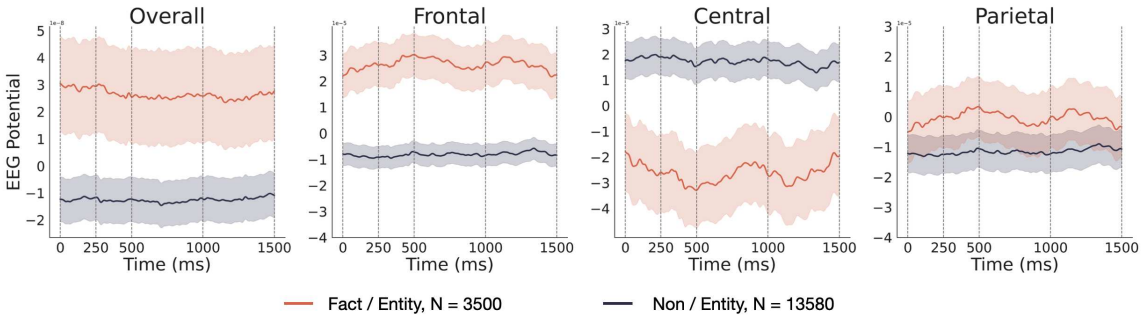

Figure 5. EEG signals obtained using the extraction method of named entity recognition.

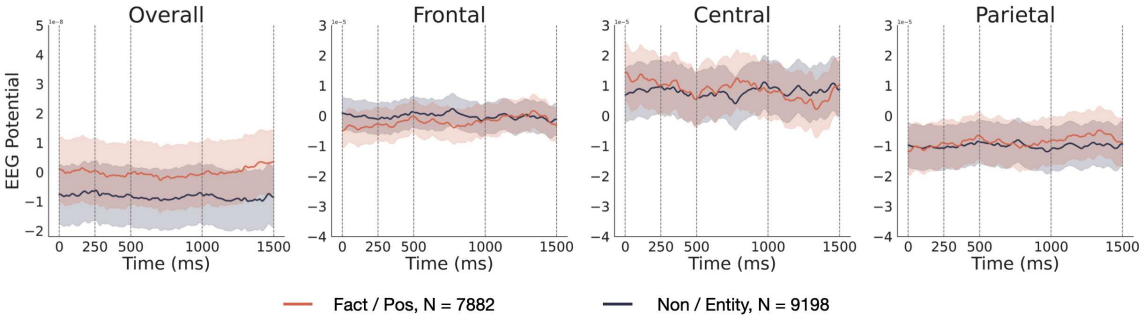

Figure 6. EEG signals obtained using the extraction method of part-of-speech tagging.

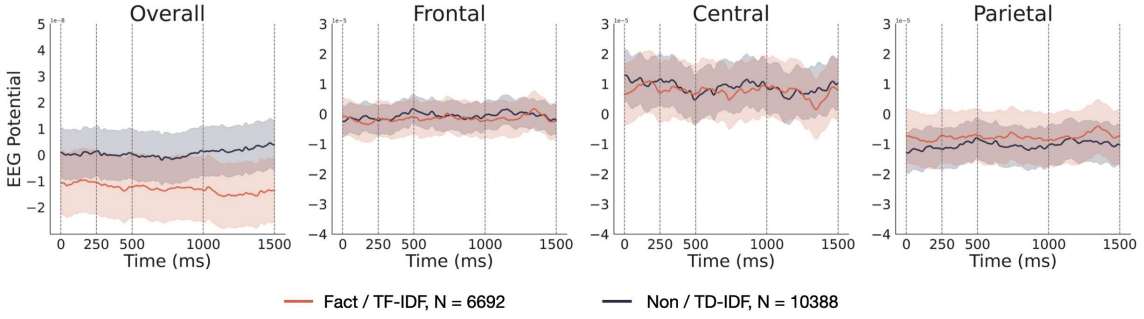

Figure 7. EEG signals obtained using the extraction method of TF-IDF.

The EEG signals corresponding to factual and non-factual words obtained using the "entity" extraction method exhibit the most significant differences, with pronounced disparities observed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

across all four AOIs (Area of Interest). This outcome aligns with previous findings based on cosine similarity, further underscoring that the distinction between groups of factual and non-factual words extracted through named entity recognition (entity) is most pronounced in the EEG signals.

4.2. Correlation between the Human Brain and Models

This research aims to examine whether the methods of extracting factual words (features), the models for generating word vectors (models), and the time windows of EEG signals (TOI) would significantly influence the representational similarity between human brain activity and model across different Area of Interest (AOI). Therefore, a three-factor repeated measures Analysis of Variance (ANOVA) can be conducted for each AOI. The three independent variables encompass features, models, and TOIs. The features can be categorized into four types: "dependency," "entity," "pos," and "TF-IDF"; the models for generating word vectors can be divided into three types: "GloVe," "BERT," and "GPT-2"; and the time windows of the EEG signals are split into three categories: "0-1500ms (Overall)", "250-500ms (N400)," and "500-1000ms (P600)." This study conducted detailed statistical analyses on the representational similarity obtained for each combi-nation, aiming to ascertain if different Features, Models, and TOIs would influence the representational similarity between the human brain and the model.

4.2.1. All 28 channels (Overall)

The descriptive statistical results of the RSA scores for the human brain and model across all channels are presented in Table A1 below. Figure 8 depicts the distribution of representational similarity between human brain activity and language models at three different TOIs for various models and fact word extraction methods using boxplots. The error bars represent the 95% confidence interval (*CI*).

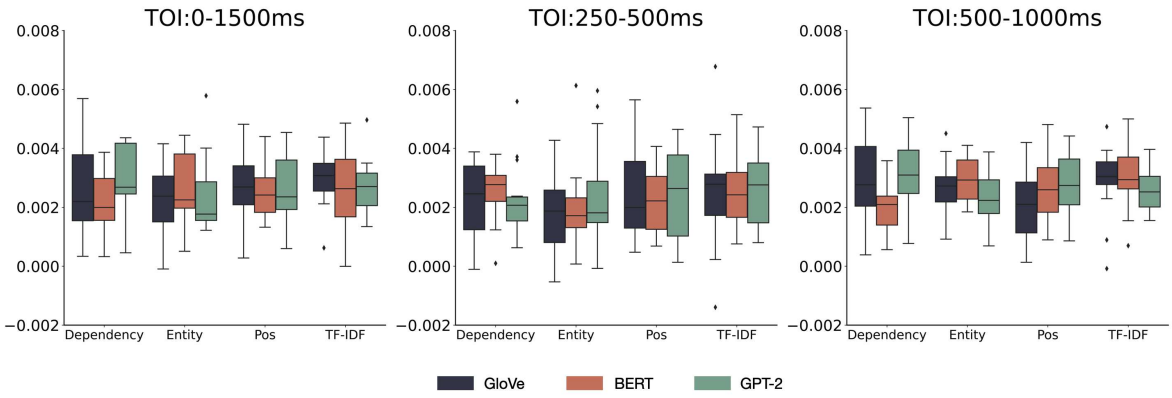

Figure 8. Distribution of RSA scores across all channels.

Before conducting a repeated measure ANOVA on the RSA scores of the 28 channels, we first performed Mauchly's Sphericity Test. We report the results of the sphericity test for the main effects and interactions that met the assumption of sphericity. For effects that violated this assumption, we applied the Greenhouse-Geisser correction. On these 28 channels, the main effects of 'Model,' 'Feature,' and 'TOI,' as well as the interaction effect between 'Model' and 'Feature,' met the assumption of sphericity. Thus, we conducted a repeated measures ANOVA using the original degrees of freedom. For other three-way interactions that did not meet the assumption, we used the Greenhouse-Geisser correction. The results of the main and interaction effects are shown in Table A5.

For the 28 channels, the main effect results are as follows: For the three models, the sphericity assumption was met ($p = 0.782$), with an ANOVA result of $F(2) = 0.334$, $p = 0.719$, indicating that the main effect of the chosen model categories on RSA scores was not significant. For the four extraction methods, the sphericity assumption was met ($p = 0.98$) with an ANOVA result of $F(3) = 0.701$, $p = 0.557$, suggesting that the main effect of the chosen fact word extraction methods on RSA scores was not

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

significant. For the times of interest, the sphericity assumption was met ($p = 0.155$) with an ANOVA result of $F(2) = 0.334$, $p = 0.719$, indicating that the main effect of the chosen periods of interest on RSA scores was not significant.

The results for the interactions are as follows: For the interaction between the model and extraction method, the sphericity assumption was met ($p = 0.970$) with an ANOVA result of $F(6) = 0.68$, $p = 0.666$, suggesting that the interaction between the two factors was not significant. For the interaction between the model and Time of Interest (TOI), the sphericity assumption was not met ($p = 0.046$), so the Greenhouse-Geisser correction was applied, resulting in $F(2.338) = 0.224$, $p = 0.833$, indicating a non-significant interaction. For the interaction between the extraction method and TOI, the sphericity assumption was not met ($p = 0.044$), so the Greenhouse-Geisser correction was applied, resulting in $F(3.112) = 0.353$, $p = 0.794$, indicating a non-significant interaction. For the three-way interaction, the sphericity assumption was not met ($p < 0.001$). After applying the Greenhouse-Geisser correction, the result was $F(4.336) = 2.253$, $p = 0.070$, suggesting that the interaction among the three factors was insignificant.

4.2.2. Ten channels located above the frontal lobe (Frontal)

Ten previously selected 28 EEG channels are located above the frontal lobe. The descriptive statistical results of the RSA scores for the brain and model based on these 10 EEG channels are presented in Table A2. Figure 9 depicts the distribution of representational similarity between human brain activity and language models at three different TOIs for various models and fact word extraction methods using boxplots. The error bars represent the 95% *CI*.

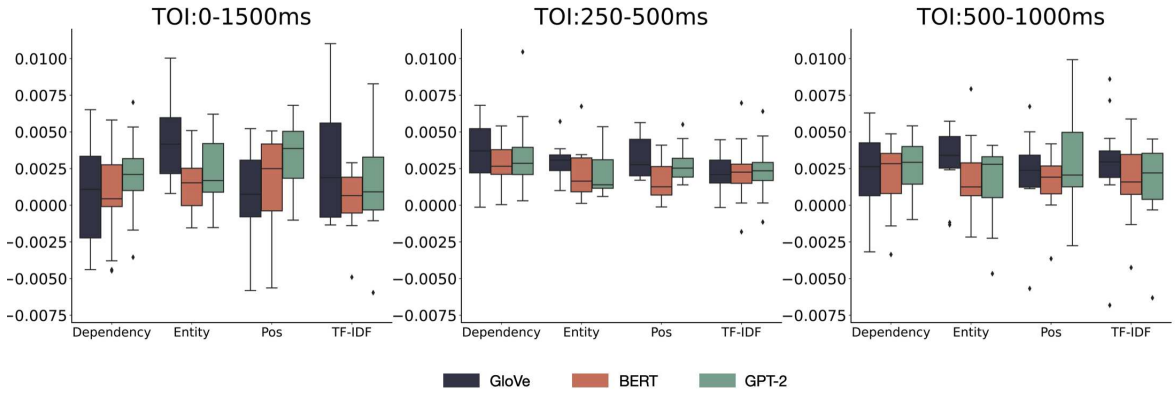

Figure 9. Distribution of RSA scores for channels located above the frontal lobe.

The results for the main and interaction effects from the repeated measures ANOVA are presented in Table A6. For the ten channels above the frontal lobe, the following are the results for the main effect analysis: For the three models, the sphericity assumption was not met ($p = 0.022$). After applying the Greenhouse-Geisser correction, the result was $F(1.36) = 5.301$, $p = 0.025$, indicating a significant main effect of the selected three model categories on the RSA scores. The sphericity assumption was met for the four extraction methods ($p = 0.865$), resulting in $F(3) = 0.851$, $p = 0.475$, indicating that the main effect of the selected fact-word extraction methods on the RSA scores was not significant. For the TOI, the sphericity assumption was met ($p = 0.647$), resulting in $F(2) = 3.706$, $p = 0.038$, indicating a significant main effect of the selected TOIs on the RSA scores.

Interaction results are as follows: For the interaction between models and extraction methods, the sphericity assumption was met ($p = 0.073$) with $F(6) = 2.139$, $p = 0.058$, in-dicating no significant interaction. For the interaction between models and TOI, the sphericity assumption was not met ($p = 0.039$), and after applying the Greenhouse-Geisser correction, the result was $F(2.418) = 0.651$, $p = 0.556$, indicating no significant interaction. The sphericity assumption was met for the interaction between the extraction method and TOI ($p = 0.507$), resulting in $F(6) = 1.519$, $p = 0.183$, indicating no significant

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

interaction. The sphericity assumption was not met for the interaction among the three factors ($p < 0.001$). After applying the Greenhouse-Geisser correction, the result was F(5.054) = 1.081, $p = 0.379$, indicating no significant interaction.

Given the significant main effects of the three models and three TOIs on the RSA scores, post-hoc tests were conducted on these two factors. The pairwise comparison results after the Bonferroni correction are presented in Table 6 and Table 7.

Table 6. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Model 1	Model 2	Difference of *M* (1-2)	*p*	95% *CI*	
				Lower	Upper
GloVe	BERT	0.001	0.050	-1.23e-06	0.002
	GPT-2	<0.001	0.384	0	0.001
BERT	GloVe	-0.001	0.050	-0.002	1.23e-06
	GPT-2	-0.001	0.226	-0.002	0
GPT-2	BERT	0.001	0.226	0	0.002
	GloVe	<0.001	0.384	-0.001	0

Table 7. This is a table caption. Tables should be placed in the main text near to the first time they are cited

TOI 1	TOI 2	Difference of *M* (1-2)	*p*	95% *CI*	
				Lower	Upper
Overall*	N400*	-0.001	0.041	-0.002	-3.00e-05
	P600	<0.001	0.772	-0.001	0.001
N400*	Overall*	0.001	0.041	3.00e-05	0.002
	P600	<0.001	0.448	0	0.001
P600	N400	<0.001	0.448	-0.001	0
	Overall	<0.001	0.772	-0.001	0.001

Overall: 0-1500ms; N400: 250-500ms; P600: 500-1000ms

From the perspective of the ten channels above the frontal lobe, although the ANOVA results indicate significant differences among the three models, the post-hoc test results suggest no significant differences between any two of the three models.

Regarding the three TOIs, post-hoc test results indicate that the RSA for the 0-1500ms interval is significantly lower than the 250-500ms interval ($p = 0.041$), but there was no significant difference compared to the 500-1000ms interval ($p = 0.772$). Additionally, there was no significant difference between the RSAs for the 250-500ms and 500-1000ms in-tervals ($p = 0.448$).

4.2.3. Ten channels located above the central region (Central)

Ten previously selected 28 EEG channels are located above the central region. The descriptive statistics for the RSA scores of the brain and the model on these 10 EEG channels are shown in Table A3. Figure 10 depicts the distribution of representational similarity between human brain activity and language models at three different TOIs for various models and fact word extraction methods using boxplots. The error bars represent the 95% CI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

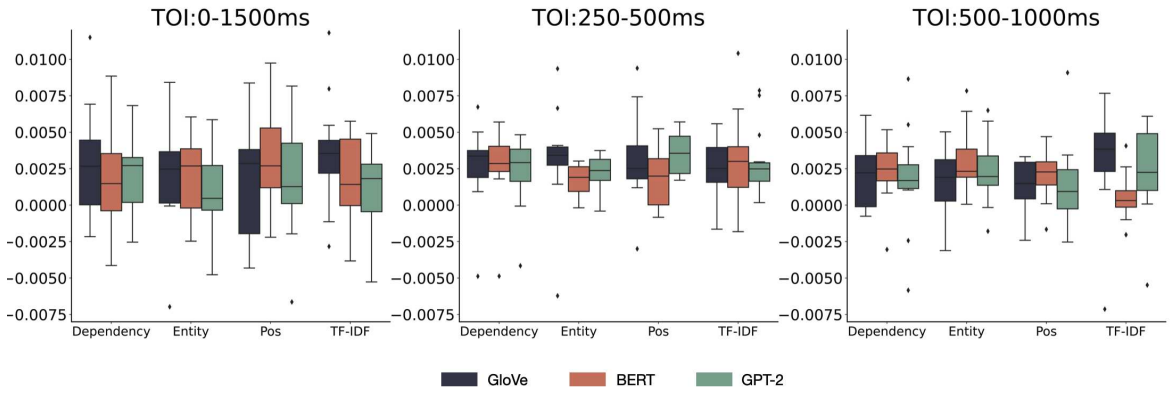

Figure 10. Distribution of RSA scores for channels located above the frontal lobe.

The results for the main and interaction effects from the repeated measures ANOVA are presented in Table A7. For the ten channels located above the central region, the results for the main effects analysis are as follows: For the three models, the sphericity assumption was met (p = 0.321), resulting in an ANOVA outcome of $F(2)$ = 0.048, p = 0.953, indicating that the main effect of the chosen model categories on RSA scores was not significant. For the four ex-traction methods, the sphericity assumption was not met (p = 0.014), so the Green-house-Geisser correction was applied, resulting in a corrected ANOVA of $F(1.709)$ = 0.166, p = 0.816, indicating that the main effect of the chosen fact word extraction methods on RSA scores was not significant. For the times of interest, the sphericity assumption was not met (p = 0.031), leading to a corrected ANOVA result of $F(1.389)$ = 4.140, p = 0.046, suggesting a significant main effect of the chosen times of interest on RSA scores.

The interaction between the model and extraction method met the sphericity as-sumption (p = 0.073), with an ANOVA result of $F(6)$ = 0.68, p = 0.773, indicating no sig-nificant interaction. For the interaction between the model and TOI, the sphericity assumption was not met (p = 0.003), so the Greenhouse-Geisser correction was applied, yielding a corrected ANOVA of $F(1.948)$ = 1.223, p = 0.310, indicating no sig-nificant interaction. The sphericity assumption was violated for the interaction between the extraction method and TOI (p = 0.005), leading to a corrected ANOVA result of $F(3.588)$ = 0.395, p = 0.791. For the three-way interaction, the sphericity assumption was not met (p = 0.020), resulting in a corrected ANOVA of $F(4.308)$ = 2.150, p = 0.082. Given the significant main effect of the times of interest on RSA scores of the brain and model, post hoc tests were conducted with Bonferroni correction, and the paired comparison results are shown in Table 8.

Table 8. This is a table caption. Tables should be placed in the main text near to the first time they are cited

TOI 1	TOI 2	Difference of *M* (1-2)	*p*	95% *CI*	
				Lower	Upper
Overall	N400	-0.001	0.165	-0.003	0
	P600	<0.001	1	-0.001	0.001
N400	Overall	0.001	0.165	0	0.003
	P600	<0.001	0.131	0	0.002
P600	N400	<0.001	0.131	-0.002	0
	Overall	<0.001	1	-0.001	0.001

Overall: 0-1500ms; N400: 250-500ms; P600: 500-1000ms

For the three TOIs, pairwise differences were not significant. However, the RSA corresponding to N400 was higher than overall and P600.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

4.2.4. Eight channels located above the parietal lobe (Parietal)

Ten previously selected 28 EEG channels are located above the parietal lobe. The descriptive statistics for the RSA scores of the brain and the model on these 8 EEG channels are shown in Table A4. Figure 11 depicts the distribution of representational similarity between human brain activity and language models at three different TOIs for various models and fact word extraction methods using boxplots. The error bars represent the 95% *CI*.

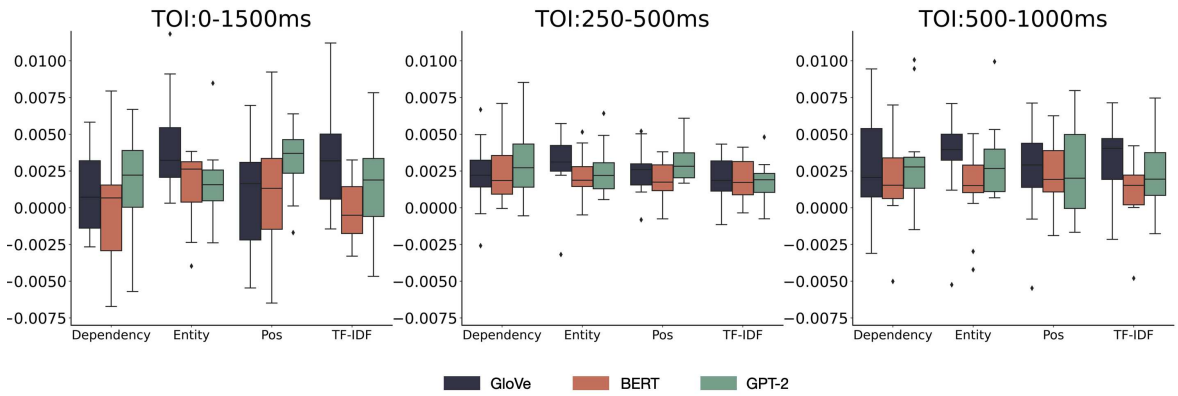

Figure 11. Distribution of RSA scores for channels located above the parietal lobe.

The results for the main and interaction effects from the repeated measures ANOVA are presented in Table A8. For the eight channels located above the parietal lobe, the main effects analysis results are as follows: For the three models, the sphericity assumption was not met (p = 0.012), leading to the application of the Greenhouse-Geisser correction, yielding a cor-rected ANOVA of $F(1.316) = 8.35$, p = 0.007, indicating a significant main effect of the model categories chosen on the RSA scores. For the four extraction methods, the sphericity assumption was not met (p = 0.021), necessitating the Greenhouse-Geisser correction, with the corrected ANOVA result being $F(1.875) = 0.415$, p = 0.652, implying the chosen fact word extraction methods did not have a significant main effect on RSA scores. The sphericity assumption was met for the times of interest (p = 0.156), resulting in an ANOVA outcome of $F(2) = 3.584$, p = 0.042, suggesting a significant main effect of the selected TOI on RSA scores.

The interaction between the model and extraction method did not meet the sphericity assumption (p = 0.033), leading to the Greenhouse-Geisser correction and a corrected ANOVA result of $F(3.468)$ = 2.012, p = 0.118, indicating no significant interaction. The interaction between the model and TOI did not meet the sphericity assumption (p = 0.027), leading to a corrected ANOVA of $F(2.609) = 1.706$, p = 0.189 after applying the Greenhouse-Geisser correction. The interaction between the extraction method and TOI did not meet the sphericity assumption (p = 0.043), resulting in a corrected ANOVA of $F(3.382) = 1.066$, p = 0.378 after the Greenhouse-Geisser correction. For the three-way interaction, the sphericity assumption was met (p = 0.051), leading to an ANOVA result of $F(12) = 1.071$, p = 0.388, implying no significant interaction.

Given the significant main effects of the three models and TOIs on the RSA scores of the brain and model, post hoc tests were conducted using the Bonferroni correction, with paired comparison results presented in Table 9 and Table 10.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

Table 9. Post-hoc Test for Main Effects of the Three Models on RSA Scores.

Model 1	Model 2	Difference of *M* (1-2)	*p*	95% *CI*	
				Lower	Upper
GloVe	BERT	0.001	0.054	-1.56e-05	0.002
	GPT-2	<0.001	1	-0.001	0
BERT	GloVe	-0.001	0.054	-0.002	1.56e-05
	GPT-2	-0.001	0.009	-0.002	0
GPT-2	BERT	0.001	0.009	0	0.002
	GloVe	<0.001	1	0	0.001

Table 10. Post-hoc Test for Main Effects of the Three TOIs on RSA Scores.

TOI 1	TOI 2	Difference of *M* (1-2)	*p*	95% *CI*	
				Lower	Upper
Overall*	N400*	-0.001	0.072	-0.001	5.22e-05
	P600	-0.001	0.191	-0.002	0
N400*	Overall*	0.001	0.072	-5.22e-05	0.001
	P600	<0.001	1	-0.001	0.001
P600	N400	<0.001	1	-0.001	0.001
	Overall	0.001	0.191	0	0.002

From the perspective of the eight channels above the parietal lobe, post hoc tests for the three models indicated that while there were no significant differences between the GloVe model and the BERT (p = 0.054) or GPT-2 models (p = 1), the RSA scores for the BERT model with the brain were significantly lower than those for the GPT-2 model with the brain (p = 0.009).

Regarding the three TOIs, pairwise differences were not significant.

5. Discussion

This study aims to investigate the factual accuracy problems of abstractive summa-rization models using EEG signals captured during natural human reading and to en-hance the interpretability of these models by analyzing the activity correlation between the brain and the model. Specifically, the study used cosine similarity and EEG signal curves to compare the brain's sensitivity to different fact extraction methods (including dependency relations, named entity recognition, part-of-speech tagging, and TF-IDF). The correlation between different language models (GloVe, BERT, GPT-2) and brain activity was examined using the representational similarity matrices of word vectors and EEG signals.

5.1. The human brain exhibits varying sensitivities to different fact extraction methods

The results indicate a significant difference in cosine similarity between fact and non-fact word groups derived from different fact extraction methods.

1. Named entity recognition yielded the lowest cosine similarity between the two groups, approaching -1, indicating high differentiation by the EEG signals.
2. Word groups extracted based on dependency relations followed closely reached -0.9, showcasing a strong negative correlation and implying a high differentiation by the EEG signals.
3. In contrast, part-of-speech tagging and TF-IDF methods did not demonstrate sig-nificant differentiation by EEG signals.

Furthermore, upon examination of the EEG signal curves, a pronounced distinction between fact and non-fact words extracted via the named entity method was evident. This observation was consistent with the conclusions derived from the cosine similarity analysis. Past research has shown

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

that the brain processes proper nouns distinctly, re-quiring a broader neural network activation and cognitive resources than common nouns. This distinction is reflected in both the N400 and P300 ERP components.[38,39] The factual words extracted using named entity recognition contain many proper nouns, contributing to the high discriminability in EEG signals between the two groups of words derived from it. Moreover, ERP components like LAN and P600 are strongly associated with de-pendency relations. The P600 component is typically related to syntactic violations, and the violation of dependency relations, such as subject-verb disagreement, also triggers the P600.[22] The LAN component usually appears after morphological or syntactic viola-tions, especially in the early stages of processing a sentence, and its amplitude typically increases when encountering mismatched dependency relations, like word order or morphological errors.[40] EEG signals are notably sensitive to recognizing dependency relations in human language tasks. This sensitivity might contribute to the pronounced distinction in EEG signals between the two groups of words derived from it. However, solely relying on part-of-speech tagging to differentiate between factual and non-factual content is overly simplistic. The mere feature of part-of-speech fails to accurately capture the semantic and syntactic roles of words within specific contexts, making it challenging to discern notable differences in EEG signals. As a statistically based extraction method, TF-IDF does not align with cognitive patterns during natural reading. Consequently, it is also challenging to manifest distinct differences in EEG signals using this method.

5.2. Model selection and TOI contribute uniquely to the correlation between the human brain and the model

In comparing the correlation between different language models and brain activity, 4 Area of Interest (AOIs) were delineated. Using EEG signals from 4 channel groups, we assessed the representational similarity between the human brain and model reading activities across 4 extraction methods, three models, and 3 TOIs. The results are presented in Table 11.

Table 11. ANOVA results of RSA scores for channels located above the frontal lobe.

Within-Subjects Effect	Overall (28)	Frontal (10)	Central (10)	Parietal (8)
Model	ns	*	ns	GPT-2>BERT**
Feature	ns	ns	ns	ns
TOI*	ns	TOI2 > TOI1*	*	*
Model × Feature	ns	ns	ns	ns
Model × TOI	ns	ns	ns	ns
Feature × TOI	ns	ns	ns	ns
Model × Feature × TOI	ns	ns	ns	ns

ns: p>.05. *: p<.05. **: p<.01. ***: p<.001; TOI_1: 0-1500ms, TOI_2: 250-500ms, TOI_3: 500-1000ms.

Cells marked with an asterisk (*) indicate significant main or interaction effects in the variance analysis, but post-hoc tests reveal no significant differences between the levels of the variables. Across the four AOIs, there were no significant interaction effects among extraction methods, model selection, and TOIs, suggesting that the impact of each variable on the RSA scores is independent, implying that each variable (e.g., extraction method, model selection, or TOI) contributes uniquely to the correlation between the human brain and the model.

Different TOIs may correspond to various linguistic processing stages within the brain. For instance, the N400 is typically associated with semantic violations or unex-pected words. The 250-500ms TOI is often utilized to study the N400 time window,[41] suggesting that this TOI likely encompasses the semantic interpretation process. Con-sequently, the higher RSA scores between the human brain and the model during the TOI encompassing the N400, compared to the scores across the entire duration, might suggest that the model, to some extent, captures information related to semantic pro-cessing. Additionally, compared to other types of information (such as syntax or back-ground knowledge), the model appears to be more sensitive to semantic data. Experi-mental results from the channels above the parietal lobe indicate that the RSA scores for GPT-2 are significantly higher than

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

those for BERT. The findings may suggest that specific processing strategies in GPT-2 resonate more with the linguistic processing modalities of this particular brain region. From the model's perspective, BERT operates bidirectionally, considering the context before and after each word. In contrast, GPT-2 processes text unidirectionally, which might more closely mirror the participants' natural reading pattern, especially given that words were presented sequentially during the experiment.

6. Conclusions

The brain is considered the only system capable of understanding language. As interest in exploring the interpretability of deep language models grows, more re-searchers are attempting to leverage the brain's linguistic cognitive mechanisms to ex-plore the "black box" of these models. Arana et al.[42] categorized the methods for studying the association between models and brain activity into three types. The first category involves recording human behavioral data, brain activity data, and model outputs during linguistic tasks and then directly comparing and analyzing the related patterns between the human brain and the model. Some researchers have employed linear models to fit the activations of GPT-2 in order to predict fMRI responses. By measuring the degree of mapping between GPT-2 and the human brain, they found that the activations in GPT-2 partially reflected the semantic representations generated by the human brain while processing spoken stories. This discovery provided compelling ev-idence for the feasibility of using cognitive neuroscience to interpret models. The second category focuses on specific derived metrics, extracting features from human brain data and model outputs for quantitative comparative analysis. One of the most commonly used metrics compares the model's "surprise" and brain activity. The concept of "surprise" originates from the field of information theory and is employed to measure the unex-pectedness of a stimulus. Heilbron et al. [42] found that the "surprise" of vocabulary in the model could effectively explain and predict brain responses. Their findings support the notion that the brain continually engages in probabilistic predictions. This approach offers a new bridge for connecting models and brain activities. The third category compares the geometric representations extracted from human behavior or brain activity with those derived from model activity. The most prevalent method used for this comparison is the RSA employed in this study. RSA does not investigate a one-to-one correspondence in word representation between the model and the human brain. Instead, it examines the overall geometric structure similarities in their representations. The application of RSA reveals which parts of a neural network model are most similar to brain neural signals and enables cross-validation between brain data and various computational models, aiding in determining which model best accounts for brain activity [34].

In subsequent research, we plan to employ various correlation methods to investigate the differences in distinct fact extraction approaches and to understand how these methods influence the similarity between models and brain activity. Existing research has cor-related different layers of the BERT model with the human language comprehension process, enhancing the model's correlation with brain activity and performance. There-fore, plans include leveraging the relationship between the model and human brain activity to refine the model's architecture, aiming to improve its performance further. Therefore, in subsequent studies, we plan to leverage the correlations between the model and human brain activity to refine the model's structure, aiming to enhance its perfor-mance in the task of ABS. To better understand this correlation, we will integrate other cognitive neuroscience research methods, such as fMRI and functional Near-Infrared Spectroscopy (fNIRS), to provide more temporal information on brain activity. Ultimately, through these investigations, we hope to identify a more precise approach to explain the relationship between the brain and deep language models, offering valuable insights for future research in cognitive neuroscience and artificial intelligence.

Author Contributions: Conceptualization, L.L. and L.Z.; methodology, Yq.Z. and Z.Z.; software, Yb.Z. and H.S.; valida-tion, Y.L., Yq.Z. and L.D.; formal analysis, Z.Z.; investigation, Yb.Z., Y.L. and L.D.; resources, L.L. and L.Z.; data curation, H.S. and S.G.; writing—original draft preparation, Z.Z.; writing—review and editing, L.L. and L.Z.;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

visualization, Z.Z.; supervision, Yq.Z.; project administration, L.L.; funding acquisition, L.L. All authors have read and agreed to the published version of the manu-script.

Funding: This research was funded by National Natural Science Foundation of China (NSFC), No. 62176024 and Engineering Research Center of Information Networks, Ministry of Education.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and ap-proved by the Ethics Committee of the Beijing University of Posts and Telecommunications.

Data Availability Statement: the data can be found here: https://drive.google.com/open?id=1Qq9P4ewmiSYDRxhUFPlHfDSZ5of6dZTt&usp=drive_fs.

Acknowledgments: We are immensely grateful to the Beijing University of Posts and Telecommunications students who participated in this work and provided us with their invaluable support.

Conflicts of Interest: The authors declare no conflict of interest. All authors have approved the manuscript and agreed with submission to this journal.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
NLU	Natural Language Understanding
ATS	Automatic Text Summarization
ABS	Abstractive (Text) Summarization
EEG	Electroencephalogram
RSA	resentational Similarity Analysis
TOI	Time of Interest
AOI	Area of Interest
pos	part-of-speech
TF-IDF	Term Frequency-Inverse Document Frequency

Appendix A. statistics of RSA scores

Appendix A.1. Descriptive statistics of RSA scores

Table A1. Descriptive statistics of RSA scores across all channels.

Features	Models	0-1500ms		250-500ms		500-1000ms		N
		M	*SD*	*M*	*SD*	*M*	*SD*	
Dependency	BERT	0.00213	0.00102	0.00256	0.00098	0.00194	0.00083	14
	GPT-2	0.00301	0.00121	0.00231	0.00126	0.00304	0.00133	14
	GloVE	0.00270	0.00163	0.00226	0.00134	0.00299	0.00147	14
Entity	BERT	0.00266	0.00124	0.00201	0.00141	0.00295	0.00078	14
	GPT-2	0.00237	0.00128	0.00244	0.00178	0.00228	0.00095	14
	GloVE	0.00228	0.00124	0.00184	0.00142	0.00269	0.00095	14
Pos	BERT	0.00253	0.00091	0.00223	0.00108	0.00262	0.00113	14
	GPT-2	0.00262	0.00109	0.00246	0.00160	0.00275	0.00114	14
	GloVE	0.00271	0.00117	0.00283	0.00218	0.00210	0.00125	14
TF-IDF	BERT	0.00271	0.00137	0.00250	0.00125	0.00309	0.00119	14
	GPT-2	0.00270	0.00095	0.00263	0.00126	0.00256	0.00073	14
	GloVE	0.00302	0.00099	0.00252	0.00189	0.00288	0.00121	14

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

Table A2. Descriptive statistics of RSA scores for channels located above the frontal lobe.

Features	Models	0-1500ms		250-500ms		500-1000ms		N
		M	*SD*	*M*	*SD*	*M*	*SD*	
Dependency	BERT	0.00048	0.00311	0.00276	0.00142	0.00208	0.00243	14
	GPT-2	0.00182	0.00279	0.00339	0.00253	0.00258	0.00197	14
	GloVE	0.00093	0.00334	0.00361	0.00211	0.00247	0.00261	14
Entity	BERT	0.00139	0.00189	0.00214	0.00173	0.00174	0.00260	14
	GPT-2	0.00224	0.00250	0.00206	0.00144	0.00165	0.00254	14
	GloVE	0.00435	0.00252	0.00297	0.00116	0.00401	0.00439	14
Pos	BERT	0.00157	0.00322	0.00163	0.00135	0.00154	0.00188	14
	GPT-2	0.00336	0.00255	0.00276	0.00116	0.00271	0.00324	14
	GloVE	0.00058	0.00307	0.00317	0.00134	0.00150	0.00416	14
TF-IDF	BERT	0.00047	0.00207	0.00234	0.00209	0.00164	0.00269	14
	GPT-2	0.00148	0.00355	0.00233	0.00184	0.00159	0.00280	14
	GloVE	0.00259	0.00386	0.00235	0.00130	0.00278	0.00346	14

Table A3. Descriptive statistics of RSA scores for channels located above the central region.

Features	Models	0-1500ms		250-500ms		500-1000ms		N
		M	*SD*	*M*	*SD*	*M*	*SD*	
Dependency	BERT	0.00091	0.00526	0.00276	0.00248	0.00133	0.00479	14
	GPT-2	0.00095	0.00512	0.00238	0.00232	0.00184	0.00335	14
	GloVE	0.00110	0.00803	0.00260	0.00266	0.00156	0.00599	14
Entity	BERT	0.00204	0.00264	0.00170	0.00105	0.00300	0.00219	14
	GPT-2	0.00064	0.00293	0.00227	0.00126	0.00148	0.00406	14
	GloVE	0.00157	0.00498	0.00316	0.00336	0.00032	0.00617	14
Pos	BERT	0.00340	0.00327	0.00188	0.00209	0.00215	0.00176	14
	GPT-2	0.00217	0.00496	0.00346	0.00140	0.00136	0.00287	14
	GloVE	0.00072	0.00696	0.00325	0.00305	0.00043	0.00459	14
TF-IDF	BERT	0.00175	0.00294	0.00302	0.00308	0.00049	0.00150	14
	GPT-2	0.00108	0.00261	0.00296	0.00231	0.00237	0.00301	14
	GloVE	0.00263	0.00543	0.00270	0.00187	0.00318	0.00351	14

Table A4. Descriptive statistics of RSA scores for channels located above the parietal lobe.

Features	Models	0-1500ms		250-500ms		500-1000ms		N
		M	*SD*	*M*	*SD*	*M*	*SD*	
Dependency	BERT	-0.00018	0.00397	0.00252	0.00211	0.00182	0.00275	14
	GPT-2	0.00154	0.00363	0.00290	0.00243	0.00295	0.00334	14
	GloVE	0.00027	0.00418	0.00226	0.00224	0.00298	0.00341	14
Entity	BERT	0.00163	0.00238	0.00216	0.00148	0.00149	0.00258	14
	GPT-2	0.00252	0.00460	0.00250	0.00167	0.00249	0.00400	14
	GloVE	0.00332	0.00490	0.00304	0.00211	0.00407	0.00397	14
Pos	BERT	0.00122	0.00397	0.00181	0.00143	0.00217	0.00217	14
	GPT-2	0.00382	0.00332	0.00308	0.00128	0.00327	0.00459	14
	GloVE	0.00071	0.00361	0.00258	0.00162	0.00192	0.00434	14
TF-IDF	BERT	-0.00034	0.00210	0.00191	0.00147	0.00123	0.00218	14
	GPT-2	0.00178	0.00334	0.00180	0.00135	0.00237	0.00247	14
	GloVE	0.00393	0.00498	0.00200	0.00155	0.00338	0.00252	14

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

Appendix A.2. ANOVA results of RSA scores

Table A5. ANOVA results of RSA scores across all channels.

Within-Subjects Effect	Mauchly's *p*	Adjustment	*df*	η^2	*F*	*p*
Model	0.782	-	2	4.6e-07	0.334	0.719
Feature	0.980	-	3	2.5e-06	0.701	0.557
TOI	0.155	-	2	3.7e-06	1.353	0.276
Model × Feature	0.970	-	6	1.6e-06	0.680	0.666
Model × TOI	0.046	Greenhouse-Geisser	2.338	2.8e-07	0.224	0.833
Feature × TOI	0.044	Greenhouse-Geisser	3.112	1.1e-06	0.353	0.794
Model × Feature × TOI	0.001	Greenhouse-Geisser	4.336	5.1e-06	2.253	0.070

*: $p<.05$. **: $p<.01$. ***: $p<.001$.

Table A6. ANOVA results of RSA scores for channels located above the frontal lobe.

Within-Subjects Effect	Mauchly's *p*	Adjustment	*df*	η^2	*F*	*p*
Model	0.022	Greenhouse-Geisser	1.36	6.04e-05	5.301	0.025
Feature	0.865	-	3	7.05e-06	0.851	0.475
TOI*	0.647	-	2	3.06e-05	3.706	0.038
Model × Feature	0.080	-	6	1.81e-05	2.139	0.058
Model × TOI	0.039	Greenhouse-Geisser	2.418	6.02e-06	0.651	0.556
Feature × TOI	0.507	-	6	1.12e-05	1.519	0.183
Model × Feature × TOI	0.006	Greenhouse-Geisser	5.054	1.35e-05	1.081	0.379

*: $p<.05$. **: $p<.01$. ***: $p<.001$.

Table A7. ANOVA results of RSA scores for channels located above the central region.

Within-Subjects Effect	Mauchly's *p*	Adjustment	*df*	η^2	*F*	*p*
Model	0.321	-	2	7.33e-07	0.048	0.953
Feature	0.014	Greenhouse-Geisser	1.709	1.35e-05	0.166	0.816
TOI*	0.031	Greenhouse-Geisser	1.389	9.34e-05	4.140	0.046
Model × Feature	0.073	-	6	1.04e-05	0.773	0.593
Model × TOI	0.003	Greenhouse-Geisser	1.948	1.72e-05	1.223	0.310
Feature × TOI	0.005	Greenhouse-Geisser	3.588	6.85e-06	0.395	0.791
Model × Feature × TOI	0.020	Greenhouse-Geisser	4.308	3.45e-05	2.150	0.082

*: $p<.05$. **: $p<.01$. ***: $p<.001$.

Table A8. ANOVA results of RSA scores for channels located above the parietal lobe.

Within-Subjects Effect	Mauchly's *p*	Adjustment	*df*	η^2	*F*	*p*
Model**	0.012	Greenhouse-Geisser	1.316	0	8.35	0.007
Feature	0.021	Greenhouse-Geisser	1.875	1.89e-05	0.415	0.652
TOI*	0.156	-	2	3.29e-05	3.584	0.042
Model × Feature	0.033	Greenhouse-Geisser	3.468	3.60e-05	2.012	0.118
Model × TOI	0.027	Greenhouse-Geisser	2.609	1.39e-05	1.706	0.189
Feature × TOI	0.043	Greenhouse-Geisser	3.382	1.92e-05	1.066	0.378
Model × Feature × TOI	0.051	-	12	6.32e-06	1.071	0.388

*: $p<.05$. **: $p<.01$. ***: $p<.001$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

References

1. Koh, H.Y.; Ju, J.; Liu, M.; Pan, S. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.* **2022**, *55*, 1–35.
2. Yang, Y.; Tan, Y.; Min, J.; Huang, Z. Automatic Text Summarization for Government News Reports Based on Multiple Features. *J. Supercomput.* **2023**, 1–17.
3. Su, M.-H.; Wu, C.-H.; Cheng, H.-T. A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2061–2072.
4. Pagnoni, A.; Balachandran, V.; Tsvetkov, Y. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6-11 June 2021; pp. 4812–4829.
5. Musil, T. Representations of Meaning in Neural Networks for NLP: A Thesis Proposal. In Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Online, 6-11 June 2021; pp. 24–31.
6. El Zini, J.; Awad, M. On the Explainability of Natural Language Processing Deep Models. *ACM Comput. Surv.* **2023**, *55*, 1–31.
7. Chen, Y.; Huang, R.; Pan, L.; Huang, R.; Zheng, Q.; Chen, P. A Controlled Attention for Nested Named Entity Recognition. *Cogn. Comput.* **2023**, *15*, 132–145.
8. Komamizu, T. Learning Interpretable Entity Representation in Linked Data. In Proceedings of the Database and Expert Systems Applications: 29th International Conference, Regensburg, Germany, 3–6 September 2018 ;Vol. 11029, pp. 153–168.
9. Ikhwantri, F.; Putra, J.; Yamada, H.; Tokunaga, T. Looking Deep in the Eyes: Investigating Interpretation Methods for Neural Models on Reading Tasks Using Human Eye-Movement Behaviour. *Inf. Process. Manag.* **2023**, *60*, 103195.
10. Lamprou, Z.; Pollick, F.; Moshfeghi, Y. Role of Punctuation in Semantic Mapping Between Brain and Transformer Models. In Proceedings of the International Conference on Machine Learning, Optimization, and Data Science, Tuscany, Italy, 18–22 September 2022; pp.458–472.
11. Li, S.; Xu, J. A Two-Step Abstractive Summarization Model with Asynchronous and Enriched-Information Decoding. *Neural Comput. Appl.* **2021**, *33*, 1159–1170.
12. Nallapati, R.; Zhai, F.; Zhou, B. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, USA, 4–9 February 2017; Vol. 31, No. 1.
13. Zhang, M.; Zhou, G.; Yu, W.; Huang, N.; Liu, W. A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, e7132226.
14. Li, H.; Zhu, J.; Zhang, J.; Zong, C.; He, X. Keywords-Guided Abstractive Sentence Summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 7–12 February 2020; Vol. 34, No. 05, pp. 8196–8203
15. Alambo, A.; Banerjee, T.; Thirunarayan, K.; Raymer, M. Entity-Driven Fact-Aware Abstractive Summarization of Biomedical Literature. In Proceedings of the 26th International Conference on Pattern Recognition (ICPR), Montréal, Canada, 21–25 August 2022; pp. 613–620.
16. Guan, S.; Jin, X.; Wang, Y.; Cheng, X. Link Prediction on N-Ary Relational Data. In Proceedings of the International World Wide Web Conference, New York, USA, 13-17 May 2019; pp. 583–593.
17. Lal, D.M.; Singh, K.P.; Tiwary, U.S. Highlighted Word Encoding for Abstractive Text Summarization. In Proceedings of the International Conference on Intelligent Human Computer Interaction (IHCI), Galway, Ireland, 12–14 December 2019; pp. 77–86.
18. Lal, D.M.; Singh, K.P.; Tiwary, U.S. ICE: Information Coverage Estimate for Automatic Evaluation Abstractive Summaries. *Expert Syst. Appl.* **2022**, *189*, 116064.
19. Zhang, M.; Zhou, G.; Yu, W.; Liu, W. FAR-ASS: Fact-Aware Reinforced Abstractive Sentence Summarization. *Inf. Process. Manag.* **2021**, *58*, 102478.
20. Yang, M.; Li, C.; Shen, Y.; Wu, Q.; Zhao, Z.; Chen, X. Hierarchical Human-Like Deep Neural Networks for Abstractive Text Summarization. *IEEE Trans Neural Netw Learn Syst.* **2021**, *32*, 2744–2757.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

21. Kutas, M.; Hillyard, S. Reading Senseless Sentences - Brain Potentials Reflect Semantic Incongruity. *Science.* **1980**, *207*, 203–205.
22. Osterhout, L.; Holcomb, P. Event-Related Brain Potentials Elicited by Syntactic Anomaly. *J. Mem. Lang.* **1992**, *31*, 785–806.
23. Ren, Y.; Xiong, D. CogAlign: Learning to Align Textual Neural Representations to Cognitive Language Processing Signals. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 1-6 August 2021; pp. 3758–3769.
24. Oseki, Y.; Asahara, M. Design of BCCWJ-EEG: Balanced Corpus with Human Electroencephalography. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 13-16 May 2020; pp. 189–194.
25. Oota, S.; Arora, J.; Agarwal, V.; Marreddy, M.; Gupta, M.; Surampudi, B. Taskonomy: Which NLP Tasks Are the Most Predictive of fMRI Brain Activity? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2022), Seattle, Washington, USA, 10–15 July 2022; pp. 3220–3237.
26. Antonello, R.; Turek, J.; Vo, V.; Huth, A. Low-Dimensional Structure in the Space of Language Representations Is Reflected in Brain Responses. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021; pp. 8332–8344.
27. Giorgi, I.; Golosio, B.; Esposito, M.; Cangelosi, A.; Masala, G. Modeling Multiple Language Learning in a Developmental Cognitive Architecture. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *13*, 922–933.
28. Luo, Y.; Xu, M.; Xiong, D. CogTaskonomy: Cognitively Inspired Task Taxonomy Is Beneficial to Transfer Learning in NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 904–920.
29. Lenci, A.; Sahlgren, M.; Jeuniaux, P.; Gyllensten, A.; Miliani, M. A Comparative Evaluation and Analysis of Three Generations of Distributional Semantic Models. *Lang. Resour. Eval.* **2022**, *56*, 1269–1313.
30. Zou, B.; Zheng, Y.; Shen, M.; Luo, Y.; Li, L.; Zhang, L. BEATS: An Open-Source, High-Precision, Multi-Channel EEG Acquisition Tool System. *IEEE Trans. Biomed. Circuits Syst.* **2022**, *56*, 1269–1313.
31. Peirce, J.W. PsychoPy - Psychophysics Software in Python. *J. Neurosci. Methods.* **2007**, *162*, 8–13.
32. Peirce, J.; Gray, J.R.; Simpson, S.; MacAskill, M.; Hoechenberger, R.; Sogo, H.; Kastman, E.; Lindelov, J.K. PsychoPy2: Experiments in Behavior Made Easy. *Behav. Res. Methods.* **2019**, *151*, 195–203.
33. Salton, G.; Wong, A.; Yang, C.S. A Vector Space Model for Automatic Indexing. *Commun. ACM* **1975**, *18*, 613–620.
34. He, T.; Boudewyn, M.A.; Kiat, J.E.; Sagae, K.; Luck, S.J. Neural Correlates of Word Representation Vectors in Natural Language Processing Models: Evidence from Representational Similarity Analysis of Event-Related Brain Potentials. *Psychophysiology.* **2022**, *59*, e13976.
35. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
36. Kenton, J. D. M. W. C., & Toutanova, L. K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2019) (Volume 1: Long and Short Papers), Minneapolis, Minnesota, USA, 2–7 June 2019; pp. 4171–4186.
37. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog.* **2019**, *1*, 9.
38. Adorni, R.; Manfredi, M.; Proverbio, A.M. Electro-Cortical Manifestations of Common vs. Proper Name Processing during Reading. *BRAIN Lang.* **2014**, *135*, 1–8.
39. Wang, L.; Zhu, Z.; Bastiaansen, M.; Hagoort, P.; Yang, Y. Recognizing the Emotional Valence of Names: An ERP Study. *BRAIN Lang.* **2013**, *125*, 118–127.
40. Kluender, R.; Kutas, M. Bridging the Gap - Evidence from Erps on the Processing of Unbounded Dependencies. *J. Cogn. Neurosci.* **1993**, *5*, 196–214.
41. Coderre, E.L.L.; Cohn, N. Individual Differences in the Neural Dynamics of Visual Narrative Comprehension: The Effects of Proficiency and Age of Acquisition. *Psychon. Bull. Rev.* **2023**, 1–15.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 November 2023 doi:10.20944/preprints202311.1717.v1

42. Heilbron, M.; Armeni, K.; Schoffelen, J.-M.; Hagoort, P.; de Lange, F.P. A Hierarchy of Linguistic Predictions during Natural Language Comprehension. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, e2201968119.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.