GeNeRTe: Generating Neural Representations from Text for Classification.

Anonymous ACL submission

Abstract

Advancements in language modelling over the last decade have significantly improved downstream tasks such as automated text classification. However, deploying such systems requires high computational resources and extensive training data. Human adults can effortlessly perform such tasks with minimal computational overhead and training data which prompts research into leveraging neurocognitive signals such as Electroencephalography (EEG). We compare Large Language Models (LLMs) and EEG features captured during natural reading for text classification. Additionally, we introduce GeNeRTe, a novel state-of-the-art synthetic EEG generative model. Using only a limited amount of data, GeNeRTe learns to produce synthetic EEG features for a sentence through a neural regressor that resolves the relationship between embeddings for a sentence and its natural EEG. From our experiments, we show that GeNeRTe can effectively synthesize EEG features for unseen test sentences with just 236 sentence-EEG training pairs. Furthermore, using synthetic EEG features significantly improves text classification performance and reduces computation time. Our results emphasize the potential of synthetic EEG features, providing a viable path to create a new type of physiological embedding with lower computing requirements and improved model performance in practical applications.

1 Introduction

Text classification serves as the foundation for many automated systems that are used every day such as categorizing medical documents, filtering harmful content, generating personalized reports, and more. Large improvements in the accuracy of automated text classification systems have become possible recently due to developments in natural language model architectures including transformers (Vaswani et al., 2017), which have led to large language models (LLMs) that are better at capturing semantic patterns in text. One caveat, however, is that pre-training LLMs requires vast amounts of training text. Even after pre-training, fine-tuning and deploying these models in production is very compute intensive and requires resources including GPUs and a high amount of RAM which are not widely available.

However, in this current era of deep learning and LLMs, it is worth remembering that an average healthy human adult can easily perform language tasks using minimal computational resources and without needing to train on vast amounts of data. Recent advancements in electroencephalography (EEG) and eye-tracking systems have enabled researchers to record brain data while performing reading tasks, which offers a wealth of information about internal representations of words and linguistic structures. For example, Ling et al. (2019) demonstrated how EEG pattern analyses can serve to decode the internal representation of visually presented words in healthy adults with word classification and image reconstruction from the EEG signal well above chance.

However, whilst research in this area has shown the power of EEG and other cognitive features, their practical application within NLP tasks has been less studied. The requirement to have humans read the text in order to generate cognitive features for classification, seems to negate the utility in most automated text-processing applications. There has been work (Hollenstein et al., 2019) combining word-level EEG features with static word embeddings. However, such word-based models are no longer competitive with LLMs which employ contextualised word embeddings. Thus, the challenge becomes how to effectively generate context-sensitive EEG features without requiring the collection of human brain data at test-time. Here, we compare EEG features recorded during natural reading with LLM embeddings in a text classification setting and explore the extent to which information contained within EEG features can be synthetically generated and then used in text classification tasks.

Our contributions are fourfold. First, we investigate and compare a Large Language Model (BERT) against EEG features collected from subjects during natural reading for the task of text classification. Second, we propose a novel state-of-the-art synthetic EEG generative model, GeNeRTe, that learns to output EEG features for a sentence by regressing between its sentence embeddings and its natural EEG. Our results show that GeNeRTe can produce good quality synthetic EEG from just 236 training sentence-EEG pairs. Third, we generate synthetic EEG for an unseen test set and compare its performance with baseline BERT, achieving significantly higher scores on the classification task. Fourth, we test our generative model on a separate benchmark dataset and demonstrate its superior performance to baseline BERT.

2 Background and Related Work

In this section, we first discuss text classification using word embedding methods (Section 2.1). We then explore what EEG features mean for language comprehension (Section 2.2). Finally, we discuss the studies that have incorporated cognitive features with NLP models and studies that have proposed models to generate synthetic cognitive features in Sections 2.3 and 2.4.

2.1 Word Embeddings for Text Classification

The foundation of modern LLMs are word embeddings. Word embeddings are n-dimensional vector representations of words which form a vector (or semantic) space with the property that words which are close together are typically similar in meaning. Static (or non-contextual) word embeddings were first popularised by Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Whilst the operational mechanisms are different, the underlying principle in both is one of distributional similarity: words are considered similar if they have similar co-occurrences. Word embeddings are trained on very large text corpora to learn meaning from context but they are non-contextual in the sense that each word in the vocabulary has a single static embedding independent of its use in a particular sentence or context. Static word embeddings have been much employed in text classification systems e.g., Wang et al. (2020).

Devlin et al. (2019) introduced Bidirectional Encoder Representation from Transformers (BERT) which is a Transformer-based encoder. BERT uses stacked encoders and a self-attention mechanism to learn contextualised embeddings, where each usage of a word has a different embedding dependent on the context. One of the reasons for the success of BERT, and subsequent LLMs, is the successful application of transfer learning. These models are pre-trained to carry out general language modelling tasks (e.g., Masked Language Modelling and Next Sentence Prediction) on vast amounts of unannotated language data. The representations learnt during pre-training are then utilised in the subsequent task-specific fine-tuning. BERT-based approaches have been hugely popular and successful in text classification, and still provide one of the best baseline models, achieving a very high accuracy (González-Carvajal and Garrido-Merchán, 2021).

2.2 Analyzing language comprehension through EEG.

Physiological processes recorded from the brain are studied using specific frequency bands or *brain oscillations/waves* named using the Greek alphabet: delta, theta, alpha, beta, and gamma. It is well-known that these oscillations vary depending on the tasks performed. For language, the brain requires attention, memory, and comprehension (syntactic and semantic). Williams et al. (2019) suggest that theta activity increases when focusing attention on a current task involving short-term memory and Bastiaansen et al. (2002) suggested an increase in theta activity during real-time language comprehension. Klimesch (2012) reports the involvement of alpha-band activity during temporal attention which is an important aspect of language comprehension. Beta-band has been associated with more complex linguistic functions such as semantic retrieval of lexicons, parsing sequences, and generating correct sentences Weiss and Mueller (2012). Lastly, Prystauka and Lewis (2019) also suggests that gamma-band activity is sensitive to semantic manipulations and factual inconsistencies.

The N400 response in the brain was discovered in the 1980s as an indicator of reading comprehension and linguistic manipulations (Holcomb, 1993). N400 is a late event-related potential that manifests around 400ms after the event as a negative peak. Grabner et al. (2007) report theta-band response around 200-600ms after a linguistic stimulus such as word presentation, including alpha, and beta-band inhibition 200-400ms after the stimulus. The increase in the gamma-band activity due to factually incorrect stimulus was also around 400-600ms after the stimulus (Prystauka and Lewis, 2019).

2.3 Augmenting NLP models with cognitive features

The Zurich cognitive corpus (ZuCo) (Hollenstein et al., 2018) dataset contains EEG and eye-tracking recordings from 12 subjects performing natural reading tasks. Hollenstein et al. (2019) proposed integrating the cognitive features from ZuCO with the neural-network based relation classification system of Rotsztejn et al. (2018). Specifically, they find that combining word-level EEG features from ZuCo dataset with word embeddings from GloVe (Pennington et al., 2014) as input to the relation classification model increases performance. They also proposed a method to generate a dictionary of word-level EEG features that can be used with datasets that do not provide any cognitive features to address the problem of not having EEG at test time.

In later work, Hollenstein et al. (2021) showed the advantage of using EEG features with contextual BERT embeddings. They first use BERT embeddings as input to a bi-LSTM model while keeping the BERT parameters trainable as the baseline. Second, they experiment with two approaches to decode the EEG features by using bi-LSTM and an inception Convolution Neural Network that uses multiple filters of different lengths to extract features from the EEG data (Szegedy et al., 2015). They then concatenate the output of the EEG decoder to the output of the BERT bi-LSTM to be passed to the classifier. This setup relies on the classifier being able to effectively learn the decoded EEG and the text representations. The complexity of this setup may increase training time and add additional parameters on top of the large language model. Moreover, their setup requires cognitive features at test time which is not suitable for other datasets and real-world use.

Ren and Xiong (2023) proposed CogAlign. They first train two individual Bi-LSTM encoders to learn task specific modalities using ZuCo cognitive features and word-level textual features from GloVe Pennington et al. (2014). Then they use a shared encoder with a modality discriminator to jointly learn cognitive and textual inputs in combination with the separate encoders. This shared encoder aligns both the representations by minimizing on the adversarial loss between them. Hence, at inference, they have the option only to use the textual input with transfer learning and obtain a joint textual-cognitive representation from the shared encoder for datasets having no cognitive features for the same task. While they show performance increases over previous methods, the use of non-contextual embeddings with the EEG features could be hampering the joint representations obtained from the shared encoder.

2.4 Generating synthetic cognitive features

Bolliger et al. (2023) proposed ScanDL, a model for generating synthetic eye-tracking data for texts. They train a diffusion model to predict gaze data on text. They convert word indices and their natural fixation sequence to embeddings and the model learns to reconstruct the natural fixation sequence using a diffusion process that introduces noise in the word index embeddings and then resolves that noise to get the original index embedding. They show improvement over previous state-of-the-art synthetic eye-tracking data generation Eyettention (Deng et al., 2023a). Deng et al. (2023b) use Eyettention to enhance BERT, rearranging the input token embeddings according to the predicted gaze fixations for a sentiment classification task.

We are not aware of any prior work generating synthetic EEG features and examining whether EEG alone is useful for downstream NLP tasks. This setup can potentially provide cognitive features closer to ground truth rather than a soft representation of those features. This would mean that LLMs won't be required during training time for downstream tasks, and we can potentially shift away from requiring huge amounts of computational resources for fine-tuning and deploying LLMs by creating EEG or physiological embeddings for various tasks. Hence, we not only address the potential issues from related work through our proposed generative model, but we also provide a novel solution to further bridge the gap between cognitive sciences and NLP. Moreover, it opens new opportunities to develop human cognition-inspired models that can co-relate better with human comprehension of language thereby creating language models that generalize like humans from a moderate amount of data.

3 Dataset and Feature Modelling

As discussed in Section 2.3, the Zurich cognitive corpus (ZuCo) (Hollenstein et al., 2018) dataset contains EEG and eye-tracking recordings from 12 subjects performing natural reading tasks. Here, we focus on the relation classification task. The sentences for the relation classification task were chosen from the Wikipedia relation extraction dataset which has 1110 paragraphs mentioning 4681 relations in 53 relation types. The authors of ZuCo selected approximately 40 sentences for each of 8 relation types (award, education, job title, political affiliation, wife, visited, nationality, and founder) and presented these to the subjects.

Subjects had to report whether a relation type was present for each sentence and there were 72 control sentences in the mix that did not have any relation type to check if there is actual comprehension of the sentences. Before starting the experiment, they had a practice round to familiarize themselves with the control and the stimuli. The instructions were presented as follows (Hollenstein et al., 2018):

AWARD; while reading the following sentences please watch out for the relation between a person or their work and the award they/it received or were nominated for.

Task instruction "Please read the following sentences. After you read each sentence, answer the question below. Press 6 when you are ready."

Example sentence: "She won a Nobel Prize in 1911"

"Does this sentence contain the **award** relation? [1] = Yes, [2] = No"

3.1 Electroencephalography (EEG)

The EEG data was recorded with a 128-channel non-invasive electrode system from Electrical Geodesics. The data was recorded with a sampling rate of 500Hz and filtered to retain wave frequency between 0.1 – 100Hz using the band pass filtering method. Out of the 128 channels, 105 were used for recording the scalp, 9 were used as electrooculography (EOG) channels for artifact removal (e.g., eye-blinks) and the rest of the channels that were placed on the neck and the face were discarded as they did not provide any essential data for processing. The artifacts were identified using MARA (Multiple Artifact Rejection Algorithm) which is an open-source plug-in for an effective supervised machine learning algorithm that analyses the EEG data so that the artifacts can be automatically rejected. Figure 3 in Appendix A shows the example of the raw EEG data and the pre-processed EEG data for a sentence.

To extract word-level EEG features, the EEG and eye-tracking data were synchronized by identifying shared events using the "EYE EEG extension" (Winkler et al., 2014) that can time lock the EEG data to the onset of eye fixations. The EEG features are spread across multiple frequency bands with each band serving its own cognitive function as discussed earlier. A total of 8 bands were identified and the data was band-pass filtered to get theta1 (t1: 4-6Hz), theta2 (t2: 6.5-8Hz), alpha1 (a1: 8.5-10Hz), alpha2 (a2: 10.5-13Hz), beta1 (b1: 13.5-18Hz), beta2 (b2: 18.5-30Hz), gamma1 (g1: 30.5-40Hz), and gamma2 (g2: 40-49.5Hz) frequencies. The final EEG features are power measures for each of these frequency bands in 105 channels that are the electrodes placed on the scalp. The power measurement of the EEG signals indicates the total activity in each frequency band per channel (Xiao et al., 2018). The authors used Hilbert transformation to estimate the amplitude and phase for each frequency band which was crucial for identifying fixation segments in eye-tracking data. For the sentence-level EEG features, they calculated the power of each frequency band over the full spectrum in 105 channels. This results in word-level EEG features: $8 \times 5 \times$ 105-dimensional vectors[1] for each fixated word; and sentence-level EEG features: $8 \times$ 105-dimensional vectors for each full sentence.

The EEG features for some sentences are not available for multiple subjects. Hence, for our model, we chose one of the highest scoring subjects with the most complete data following Hollenstein et al. (2019) who suggested that single-subject models performed slightly better than taking average across subjects. We first take the mean of 8×105 dimensional sentence EEG features. Each frequency band reflects a comprehension state of language as discussed in Section 2.2. Hence, this mean provides the full comprehension data from all the frequency bands allowing us to compile a final 105-dimensional sentence EEG. We then pre-process the sentences by lowercasing and removing punctuations and normalize the final sentence EEG features between 0-1 scaled by 1000x. The final dataset consists of normalized sentences, mean of the sentence EEG features in 8 frequency bands, and one of the 8 relation types of each sentence. Finally, we divide the dataset into training and testing set with 80-20 split and make sure that the test data does not contain any repetition of training data.

3.2 EEG Features Across Subjects and Relation Types:

The task specific reading experiment setup in ZuCo is particularly interesting as the authors instruct the subjects to look for a specific relation type in a sentence while recording EEG data. Wehbe et al. (2014a) recorded fMRI data from subjects while they read stories. They showed how neural representations can have distinct signatures for different stories that can be classified with high accuracy. Following this, we speculate that since the subjects were shown the relation type to look for before the sentence was presented, the EEG features recorded when reading that sentence amplified the signature for its relation type which might lead to distinct patterns of EEG features.

[1] #frequency bands = 8, #gaze features = 5

To test the above hypothesis, we create a function that groups the final sentence EEGs into their respective relation types resulting in 8 groups of sentence EEGs. We then take the mean of the sentence EEGs within each group to get the average sentence EEG features per relation type. Now, we can compare if the mean EEG features for each relation type show a unique pattern. We perform an Analysis of Variance (ANOVA) test to check if the mean EEG features across the different relation types are significantly different. Figure 1 shows the average EEG features per relation type for the best subject with the most complete data.

We can observe from Figure 1 that the EEG features for different relation types are clearly unique. This is further corroborated through the ANOVA test with the p-values well below 0.05 affirming that EEG patterns vary significantly across different relation types. This behavior is replicated for other subjects (See Figure 4 in Appendix B). Wehbe et al. (2014b) showed how word embeddings produced by recurrent neural networks can be aligned with word level brain activity recorded via Magnetoencephalography (MEG) while subjects read stories. Hence, in theory, these well separated sentence EEG features should lay a strong foundation for classification and regression modelling. The models should be able to efficiently learn the relationships between the sentence EEG features, sentence embeddings, and their relation types by aligning their unique patterns.

4 Synthetic EEG features

We propose GeNeRTe, an Encoder-Regressive Generator model that can produce synthetic EEG features for any sentence. We employ this model along with a random forest classifier for text classification. Figure 2 shows the model architecture.

Encoder-Regressive Generator: The core of our model is a deep neural regressor. The regressor outputs a 105-dimensional vector in line with the natural EEG features provided in ZuCo. The training process consists of two parts. First, we use a language model (BERT-base[2] in this case) as the encoder to extract word embeddings for the ZuCo sentences and create a lookup table where we store these embeddings as static objects to be used later for training the model. The sentence embedding is the mean of the last hidden state of BERT for all tokens. Then, the model trains to resolve the relationship between the sentence and its natural EEG by taking as input the static sentence embeddings and generating the EEG features. During training, the regressor backpropagates over the neural network and adjusts its parameters to essentially discern patterns of similarity between the sentence and its natural EEG. Our model consists of three hidden layers containing dropouts to randomly set a fraction of hidden layer nodes to 0 and batch-normalization to prevent overfitting. The forward pass uses ReLU activation. Given an input sentence embedding vector $X \in \mathbb{R}^I$ where I is the dimension of the embedding, the synthetic EEG can be generated using F(X):

$$F(X) = L_4 \circ D_3 \circ \delta \circ BN_3 \circ L_3 \circ D_2 \circ \delta \circ BN_2 \circ L_2 \circ D_1 \circ \delta \circ BN_1 \circ L_1(X)$$

where $\circ$ denotes composition of functions, L_i, D_i, BN_i for $i = 1, 2, 3, 4$ denote fully connected layers, dropout, and batch normalization respectively and δ denotes the ReLU activation.

GeNeRTe can produce synthetic EEG features for sentences resembling ZuCo relation extraction classes, removing the need for natural EEG during testing for real-world use. To test if it can maintain the same distinct patterns for relation classes, we performed the same ANOVA analysis as discussed in Section 3.2. Please refer to Figure 5 and 6 in Appendix B).

Classifier: We use a Random Forest classifier along with the synthetic EEG generator to implement the relation classification task. We first setup a parameter distribution dictionary that covers various permutations of hyperparameters, then we use RandomSearchCV with 5-fold cross validation to find the best performing parameters. The classifier takes EEG features as input and classifies it to a relation class. Hence, this setup uses only EEG data for text classification thereby eliminating the need for fine-tuning LLMs and word embeddings.

5 Experiments

There are a total of 236 training samples and 60 test samples in the ZuCo dataset that we use in this research. Each sentence is assigned one of the 8 relation classes.

Relation classification baseline: We fine tune pre-trained BERT-base on the ZuCo dataset as the baseline. We initialize a Trainer class from the transformers library[3] with the Adam optimizer and use Cross-Entropy Loss to train the model for 15 epochs. We keep BERT parameters trainable so that the model can learn to predict the relation class given the sentence and update itself.

[2]Other 'larger' flavours of BERT can potentially further increase the performance.

[3]https://huggingface.co/docs/transformers

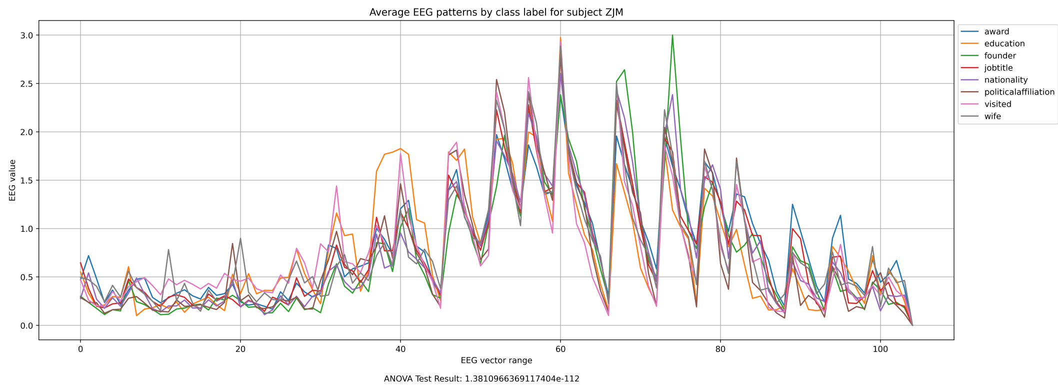

Figure 1: **Average EEG features per relation type:** The graphs illustrate group mean EEG features across eight relation types. The Y-axis denotes the EEG values. X-axis denotes the vector range (105-dimensions). ANOVA test results show p-values for the subject below the X-axis. The P-value for the subject is below 0.05 indicating statistical significance.

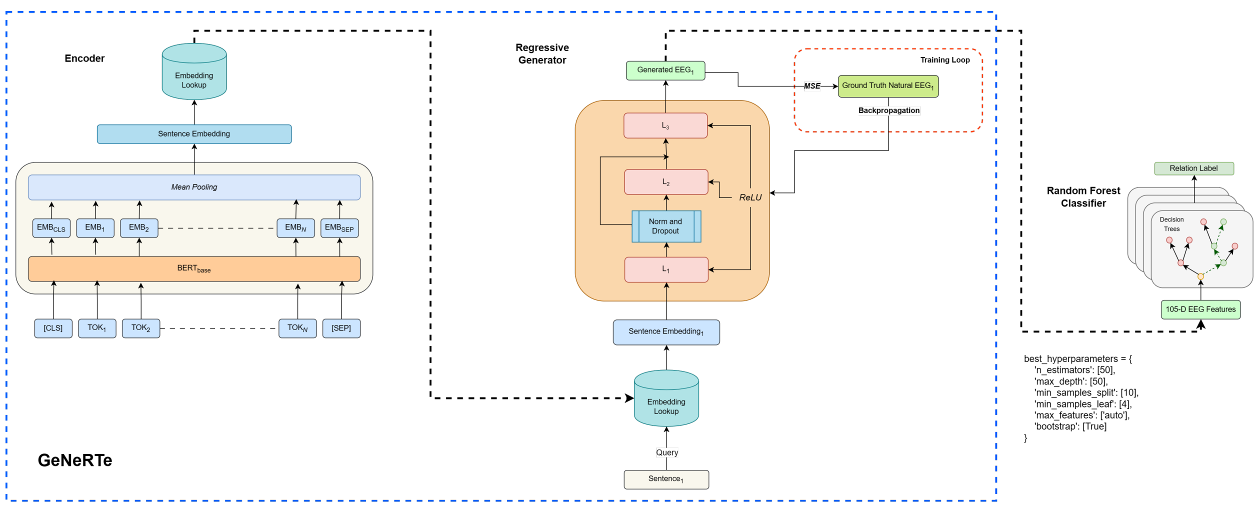

Figure 2: **GeNeRTe and Classifier Architecture:** GeNeRTe is shown within the blue box consisting of the Encoder-Regressive Generator. Encoder outputs are stored in the embedding lookup table and queried during regressor training and generation. Input EEG features for the random classifier training can be natural EEG or synthetic EEG depending on the experiment setup.

Relation classification GeNeRTe-Classifier: For our proposed model, we setup the experiments in two phases. First, we train our Encoder-Regressive Generator model on the ZuCo training set. We use Mean Squared Error Loss and a custom loss function combining Cosine Loss and MSE along with Adam optimizer and train the model for 300 epochs. We then generate synthetic EEG features for the unseen ZuCo test set. In the second phase, we use the random forest classifier with the best hyperparameters given by the RandomSearchCV process for the relation classification task. The classifier takes in EEG features as input and classifies those features into a relation class. We run the following experiments:

1. **Training and testing with natural EEG:** As the ZuCo dataset contains natural EEG for the sentences, we have the availability of EEG features at test time. Hence, we can train and test the random forest classifier with natural EEG features to evaluate the performance standard.

2. **Training with natural EEG and testing with synthetic EEG:** This test informs us about the generalization capabilities of the generative model and the nature of the synthetic data. The synthetic EEG features for the test

sentences should maintain the general direction and magnitude of its relation type learned from the training set.

3. **Training and testing with synthetic EEG:** This test is crucial to eliminate the requirement of natural EEG completely. This test also informs us about the compatibility of synthetic EEG features with itself when used for both training and testing.

Wikipedia Benchmark: The final test is to check if the generative model can perform well on a completely new dataset. We compile a benchmark dataset with sentences and their relation classes from Culotta et al. (2006). We use the same relations classes as the ZuCo dataset. There are a total of 900 training samples in the benchmark training and 225 testing samples in the benchmark testing dataset. Both the training and testing samples do not have any EEG data associated with it hence, we use our generative model to generate synthetic EEG features for both sets. First, we fine-tune BERT-base on the benchmark training set. The input to the BERT model is the sentence and the model classifies the sentences into their relation classes and updates its parameters based on the prediction error. Then we use the fine-tuned BERT model to predict the unseen test set and report the performance metrics and computation costs. We compare this to our generative model by training our random forest classifier with the synthetic EEG on the training set and making predictions on the unseen synthetic EEG features on the test set and report the same performance metrics and computation costs.

6 Results

Computation Costs: One of the prime foci of this research is to reduce the computation cost of training and setting up inference on LLMs. Fast and efficient training of LLMs requires GPUs that might not be accessible to everyone. Even with GPUs, the computation time is not modest. Our proposed model avoids the computation costs of BERT. Even combining the training and generation time of the synthetic EEG model (which needs to be trained only once) and the train and test random forest classifier is significantly faster than fine-tuning BERT on the same GPU. Table 1 shows the computation times for the experiments. It can be observed from Table 1 that the complete process of training the generator, generating the synthetic EEG, and training-testing (5-fold) on the random forest classifier is 56.57 seconds whereas fine-tuning and testing using BERT (5-fold) is 23.2 minutes.

Task	**Duration (seconds)**	
	ZuCo	Benchmark
GeNeRTe		
- Generator Training	55	55
- Generate synthetic EEG	0.5	1.07
Random Forest Classifier (RF)		
- Train-Test	0.98	3.24
BERT		
- Train/Test	1392	6540

Table 1: Computation time for GeNeRTe-Random Forest Classifier vs Fine-tuning BERT on a Tesla T4 GPU.

In addition to the significantly lower computational costs, our proposed model achieves remarkable performance increase over baseline BERT. We have three baselines. First is utilizing BERT embeddings in a neural network, the second, applying BERT embeddings in the same RF model as GeNeRTe and the third, utilizing TF-IDF vectors again in the RF model as GeNeRTe for consistency. Table 2 shows the performance for the experiments. Training and testing on natural EEG features provided in ZuCo gives an F1-Score of 94.36%. This shows the raw capability of the natural sentence EEG features. Our proposed model is able to maintain the general characteristics of the natural EEG data very well which is evident from the second experiment i.e. training on natural EEG and testing on synthetic EEG. This informs the generalization capability of our regressive model that can correlate well between word embeddings and natural EEG. Results for the third experiment shows that the synthetic EEG features are compatible with itself when used in both training and testing sets. This again shows that the regressive generator does a great job of maintaining the patterns of the EEG data. We also note that our model performs just 4.5% shy of the human subject. On the other hand, low performance of BERT classification might stem from sentences in the dataset which could point to multiple relation types.

It is evident from the above experiments that using only EEG for text classification outperforms LLMs like BERT and the benchmarking experiment further corroborates those findings. The benchmark dataset is a separate and new dataset with no EEG features associated with it. We generated synthetic EEG features for both training and testing sets and the classifier replicates the success of our third ZuCo based experiment which is training and testing with synthetic EEG data. Table 3 shows the results of the benchmark dataset. We ran the benchmark experiment for 5-folds for both BERT and the Random Forest Classifier. Our model shows a notable in-

Model	Accuracy	Precision	Recall	F1
Human Subject (ZJM)	0.9656	-	-	-
Baseline BERT	0.7266	0.7435	0.7266	0.7181
Baseline BERT Random Forest	0.7336	0.7319	0.7336	0.7190
Baseline TF-IDF Random Forest	0.5636	0.5361	0.5636	0.5184
Hollenstein et al. (2019)	-	0.683	0.648	0.651
Ren & Xiong (2023)	-	77.94	82.60	78.66
Random Forest (Train/Test on Natural EEG)	0.9436	0.9459	0.9436	0.9438
GeNeRTe Random Forest (Train on Natural/Test on Synthetic EEG)	0.9373	0.9388	0.9373	0.9371
GeNeRTe Random Forest (Train/Test on Synthetic EEG)	**0.9191**	**0.9277**	**0.9191**	**0.9207**

Table 2: Experiment results (5-fold) for task-specific relation classification. Input to the RF classifier are Natural EEG features and synthetic EEG features generated by the Encoder-Regressive model. We report Accuracy, Precision, Recall, and F1.

Model	Accuracy	Precision	Recall	F1
Baseline BERT	0.7306	0.7133	0.7306	0.7188
Baseline BERT Random Forest	0.7378	0.7291	0.7378	0.7253
Baseline TF-IDF Random Forest	0.5200	0.5265	0.5200	0.4467
GeNeRTe Random Forest (Train/Test on Synthetic EEG)	**0.7511**	**0.7453**	**0.7511**	**0.7440**

Table 3: Benchmark result comparison (5-fold) BERT vs GeNeRTe-RFClassifier. Input to the RF classifier are synthetic features generated by the Encoder-Regressive model.

crease in performance by 2-3% over baseline BERT.

7 Discussions and Conclusions

With this paper, we address our research goals to develop models that may generalize like humans with low data and computational resources using cognitive features. We investigated whether EEG only could perform better than LLMs for downstream NLP tasks. For this, we proposed a novel Encoder-Regressive generator model GeNeRTe which produces synthetic EEG data for a given sentence and we compare it with BERT in a text classification task. We trained GeNeRTe on the task specific reading dataset from ZuCo and performed three experiments to determine performance standard, generalization capability, and self-compatibility of the synthetic EEG features produced by the model. We chose text classification for the experiment setup in ZuCo hypothesizing that unique patterns could emerge for each relation type when subjects read the sentences. Text classification systems that use BERT models are deployed in many real-world systems which is why it is a strong baseline. Despite pre-training on a huge dataset, BERT performs poorly as one sentence can relate to multiple classes which is accounted for in the natural EEG features. Our model performance significantly exceeded the previous methods and the baseline by up to 30% on the unseen ZuCo test set. Furthermore, our model surpassed the baseline by 3% on the benchmark dataset (without any cognitive features) achieving overall state-of-the-art performance.

All our findings support the hypothesis that EEG features alone are beneficial for downstream NLP tasks. Specifically, the experiment setup of task-specific reading in ZuCo produces EEG features that generalize across participants in terms of emergent comprehension patterns for relation classes. It is important to note that LLMs like BERT encode the semantic and syntactic patterns of language quite well, which is why our Regressive generator was able to learn the relationship between the word embeddings and its natural EEG. Most importantly, the synthetic features produced by our generative model naturally exhibit the same comprehension patterns for relation classes as its human subject, which is very interesting. While there is no requirement for word embeddings during classification, they are important to train GeNeRTe. This setup could lead to new physiological embeddings with low computational needs, performing close to human subjects. Our research addresses the major drawback of requiring EEG features at test time and in real-world use, as it is not possible to conduct EEG data collection experiments at scale. With just a few samples, our model generates good quality synthetic features which can generalize over similar datasets as evident from our benchmark results. Our research opens new possibilities to model patterns of brain activity with applications in NLP. It is our hope that with further research, physiological embeddings could replace word embeddings in many tasks including syntactic and semantic analysis, sentence similarity, paraphrasing, and summarization.

8 Limitations

While our research shows promising results for the application of synthetic EEG in NLP, there are certain limitations. First, without any changes, our method for text classification might not be entirely applicable to other NLP problems. Second, our model is specific to relation classification experiment of ZuCo since only that experiment setup provided the distinct EEG signatures necessary to build a good synthetic EEG generator. Third, because our model depends on task-specific datasets the experimentation strategies described in relation-classificaiton experiment in ZuCo must be used to obtain comparable findings for other downstream tasks. Fourth, even though our model only needs a small amount of training data, scalability is uncertain because not everyone has access to EEG recording technology. Ultimately, additional validation and analysis of artificial EEG features is required for NLP through various downstream tasks. We acknowledge that our work is prohibitive in the sense that similar experiments to ZuCo needs to be conducted to replicate our findings in other downstream tasks and datasets. However, the results are encouraging and motivates the use case of building a database of domain-expert EEG recordings for downstream tasks. We hope to address some of these limitations in our future work, as they are critical to improving the practical usability of EEG-based NLP models.

References

M. Bastiaansen, J. van Berkum, and P. Hagoort. 2002. Event-related theta power increases in the human eeg during online sentence processing. *Neuroscience Letters*, 323(1):13–16.

L. Bolliger, D. Reich, P. Haller, D. Jakobi, P. Prasse, and L. Jäger. 2023. Scandl: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of EMNLP 2023*, page 15513.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303, New York City, USA. Association for Computational Linguistics.

S. Deng, P. Prasse, D. Reich, T. Scheffer, and L. Jäger. 2023b. Pre-trained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of EMNLP 2023*.

S. Deng, D. R. Reich, P. Prasse, P. Haller, T. Scheffer, and L. A. Jäger. 2023a. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Google, and Artificial Language. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

S. González-Carvajal and E. C. Garrido-Merchán. 2021. Comparing BERT against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

R. Grabner, C. Brunner, R. Leeb, C. Neuper, and G. Pfurtscheller. 2007. Event-related eeg theta and alpha band oscillatory responses during language translation. *Brain Research Bulletin*, 72(1):57–65.

P. J. Holcomb. 1993. Semantic priming and stimulus degradation: implications for the role of the n400 in language processing. *Psychophysiology*, 30(1):47–61.

N. Hollenstein, M. Barrett, M. Troendle, F. Bigiolli, N. Langer, and Ce. Zhang. 2019. Advancing nlp with cognitive language processing signals.

N. Hollenstein, C. Renggli, B. Glaus, M. Barrett, M. Troendle, N. Langer, and C. Zhang. 2021. Decoding eeg brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, 15.

N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1).

W. Klimesch. 2012. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16:606–617.

S. Ling, A. Lee, B. Armstrong, and A. Nestor. 2019. How are visual words represented? insights from eeg-based visual word decoding, feature derivation and image reconstruction. *Human Brain Mapping*, 40(17):5056–5068.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Y. Prystauka and A. Lewis. 2019. The power of neural oscillations to inform sentence comprehension: A linguistic perspective. *Language And Linguistics Compass*, 13(9).

Y. Ren and D. Xiong. 2023. Cogalign: Learning to align textual neural representations to cognitive language processing signals.

J. Rotsztejn, N. Hollenstein, and C. Zhang. 2018. Eth-ds3lab at semeval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 689–696.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and et al. 2015. "going deeper with convolutions". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chen Wang, Paul Nulty, and David Lillis. 2020. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Ananth Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, 9(11):e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *ACLWeb; Association for Computational Linguistics*, pages 51–62.

S. Weiss and H. Mueller. 2012. "too many betas do not spoil the broth": The role of beta brain oscillations in language processing. *Frontiers In Psychology*, 3.

C. Williams, M. Kappen, C. Hassall, B. Wright, and O. Krigolson. 2019. Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *Neuroimage*, 189:574–580.

Irene Winkler, Stephanie Brandl, Felix Horn, Eric Waldburger, Carsten Allefeld, and Michael Tangermann. 2014. Robust artifactual independent component classification for bci practitioners. *Journal of Neural Engineering*, 11(3):035013.

R. Xiao, J. Shida-Tokeshi, D. Vanderbilt, and B. Smith. 2018. Electroencephalography power and coherence changes with age and motor skill development across the first half year of life. *PLOS ONE*, 13(1).

A Appendix A

B Appendix B

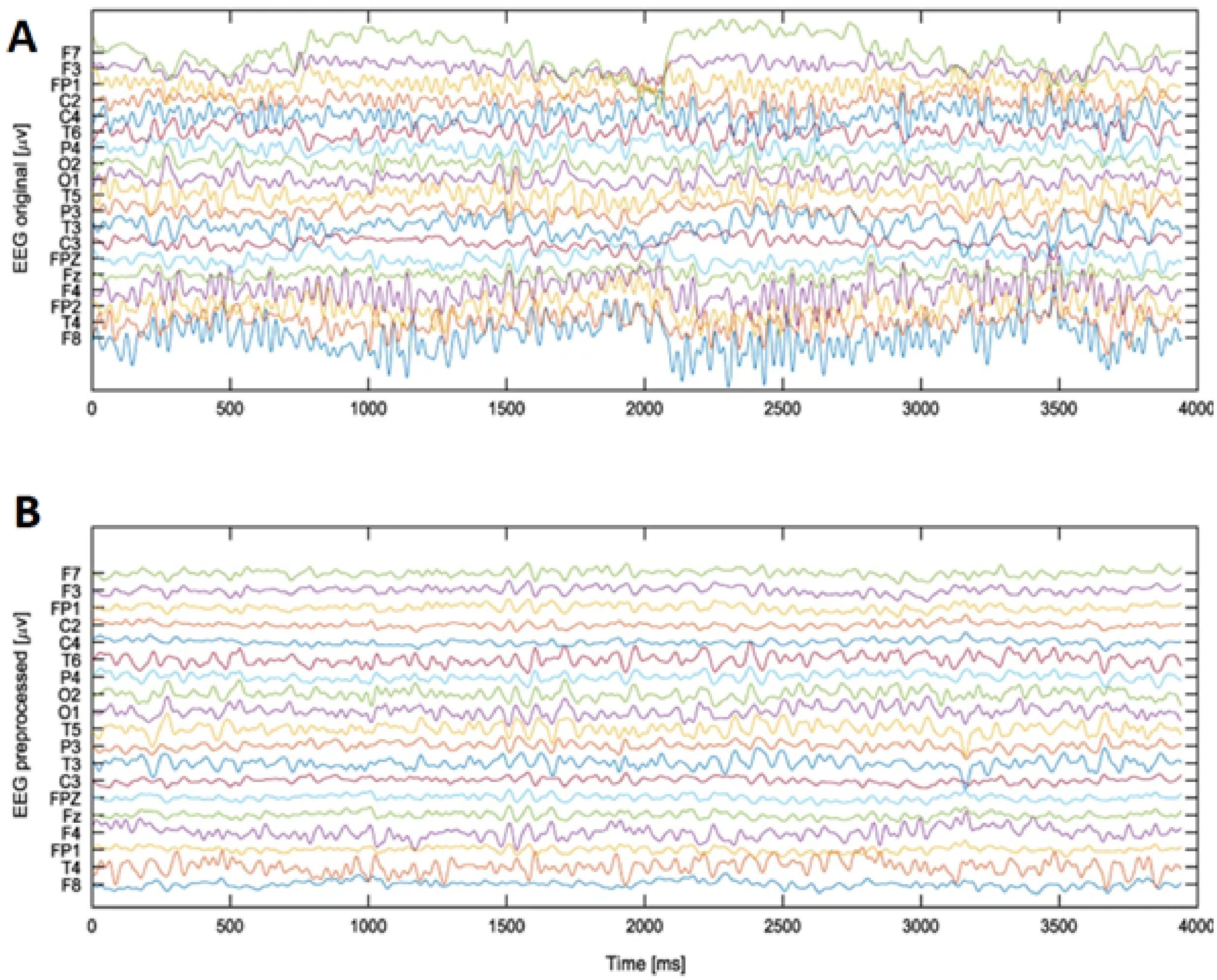

Figure 3: **Raw and Pre-processed EEG data for an example sentence (Hollenstein et al., 2018):** (A) shows the raw EEG data, (B) shows the pre-processed EEG data. The y-axis shows the brain regions where electrodes were placed. F=frontal, FP=pre-frontal, C=central, T=temporal, P=parietal, O=occipital. The x-axis shows the time.

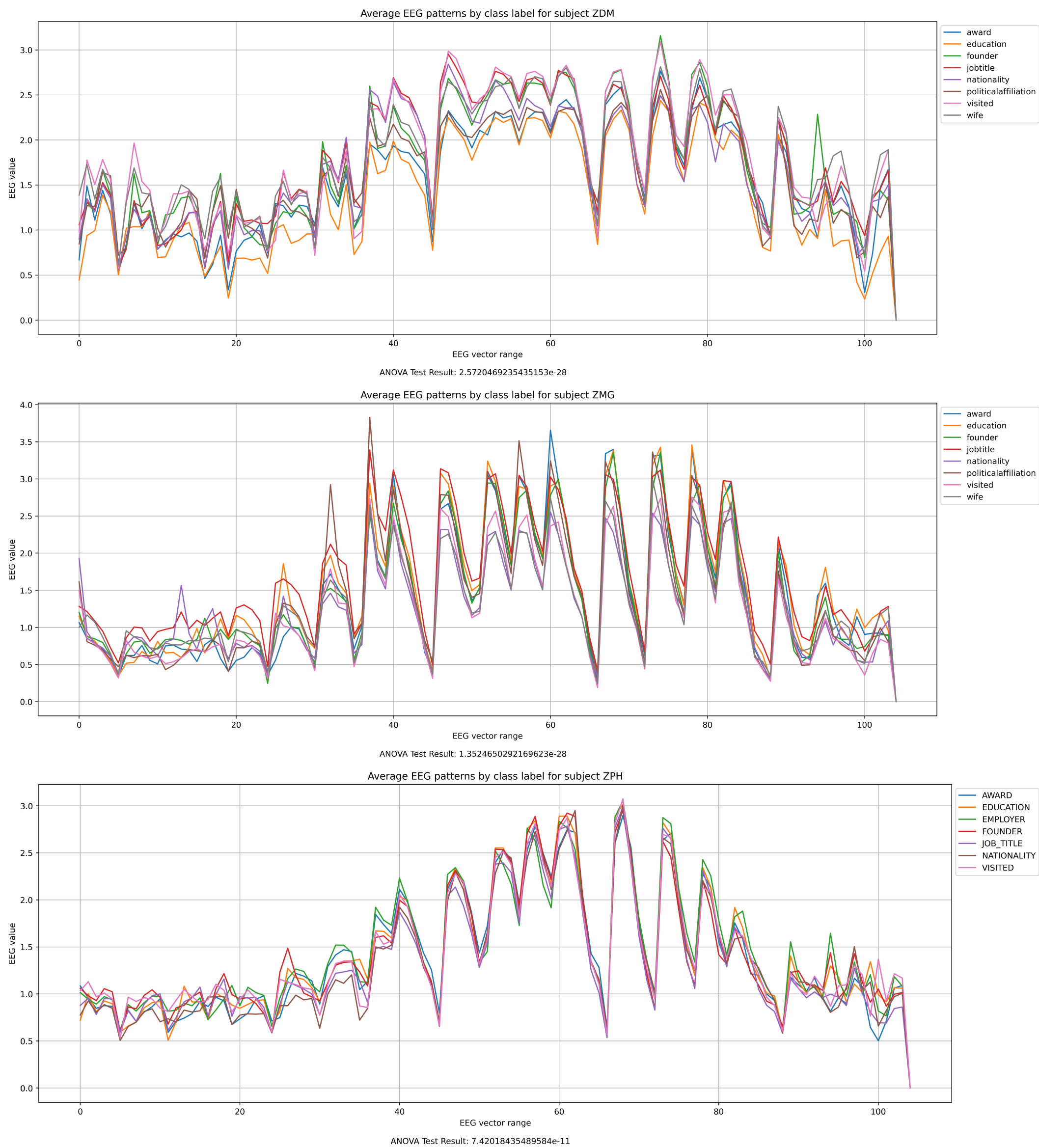

Figure 4: **Average EEG features per relation type:** The graphs illustrate group mean EEG features across eight relation types. Y-axis denotes the EEG values. X-axis denotes the vector range (105-dimensions). ANOVA test results show p-values for each subject below the X-axis. P-value for each subject is below 0.05 indicating statistical significance.

Synthetic EEG features generated by our model also show distinct patterns for each relation type which is corroborated by the ANOVA test giving a statistically significant result.

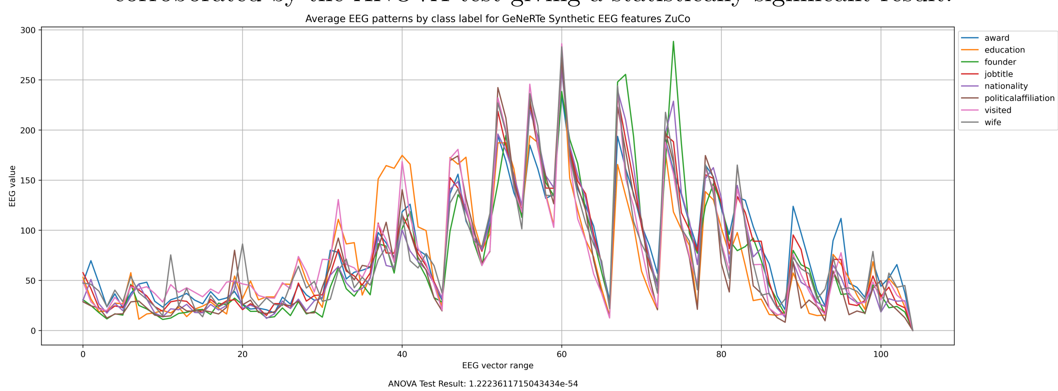

Figure 5: **Average EEG features per relation type:** The graph illustrates group mean EEG features across eight relation types as prodced by GeNeRTe for the ZuCo dataset. Y-axis denotes the EEG values. X-axis denotes the vector range (105-dimensions). ANOVA test results show p-values for each subject below the X-axis. P-value for each subject is below 0.05 indicating statistical significance.

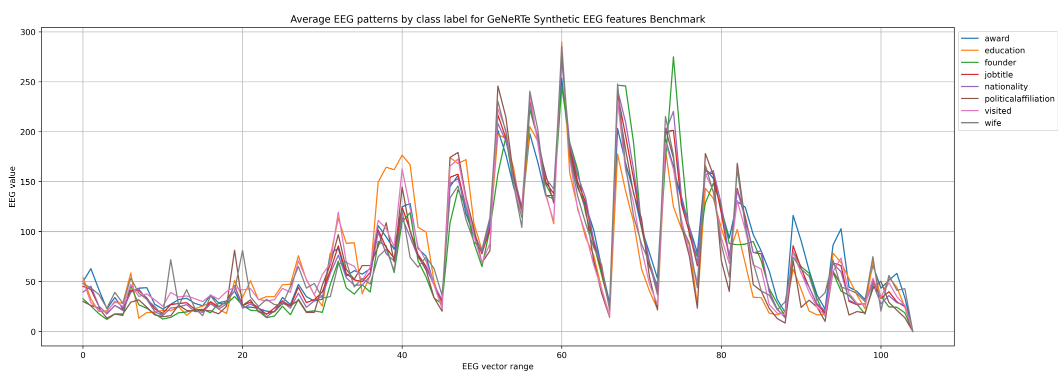

Figure 6: **Average EEG features per relation type:** The graph illustrates group mean EEG features across eight relation types as prodced by GeNeRTe for the Benchmark dataset. Y-axis denotes the EEG values. X-axis denotes the vector range (105-dimensions). ANOVA test results show p-values for each subject below the X-axis. P-value for each subject is below 0.05 indicating statistical significance.