

A Proof-of-Concept Prompt Study of User-Controlled Disclosure in LLM Generation for Text-Based BCI's for Emotional Communication

Abner Chang
abner.chang0812@gmail.com

ABSTRACT

Background/Objective: Emotional expression remains a significant challenge in text-based brain-computer interfaces (BCIs), which often struggle to convey the nuances of interpersonal communication. This study proposes a seemingly new pipeline that integrates affective BCIs (aBCIs), large language models (LLMs), and user-controlled communication settings. The study was originally intended to prove how the pipeline could improve emotional expression while preserving user autonomy. Instead, other implications were found. This does not include real BCIs, but is instead only a prompt study, examining if the concept works in theory.

Methods: A prompt-based simulation was developed to model a communication system in which hypothetical valence and arousal information, user-selected tone, and disclosure level were provided to a local LLM. All aBCI and P300-related inputs were manually simulated. A comparison was conducted between a baseline model with information regarding emotional state and communication tone, and an enhanced model containing the same information plus a user-controlled disclosure parameter of 0%, 50%, or 100%. Generated sentences were evaluated using XLM-RoBERTa sentiment classification scores, with mean, standard deviation, and Hedges' g used to characterize the resulting score distributions.

Results: The addition of the disclosure slider produced changes and variance in the sentiment classification scores. The amount and direction of these changes differed across emotional states and communication tones. Formal conditions generally produced more concentrated classifier scores, while humorous conditions showed greater variability in several emotional categories. These findings demonstrate that adding a user-controlled disclosure parameter can change the characteristics of the LLM-generated messages under otherwise same prompting conditions.

Conclusion: This proof-of-concept study demonstrates the potential for user-controlled disclosure to provide an additional level to control emotional expression in LLM-BCI communication, that it can measurably influence generated sentences, and to propose a pipeline to possibly express emotions consistent with what BCI users want.

Keywords: Large Language Model, Emotional Expression, affective Brain Computer Interface, Brain Computer Interface, Emotion Recognition, Machine Learning

INTRODUCTION

BACKGROUND AND CONTEXT

Communication is heavily influenced and facilitated by non-verbal signals, not just the words people speak.¹ It is because of this that BCIs serve as a huge potential technological advancement for the well-being of humans, for those with severe motor or speech impairments². However, how can non-verbal signals (like urgency or emotions) be better integrated and expressed into text-based Brain Computer Interfaces for people who can't speak? Severe communication impairment can substantially restrict social interaction and quality of life, creating an important motivation for assistive communication technologies³.

This is where Brain Computer Interfaces come in, and in order to mediate effective communication for people who rely on BCIs. Therefore it would be beneficial for BCIs to express emotion accurately when people are using BCIs for interpersonal communication as to not sound robotic and lose their personal voice. While text-based and voice-based BCIs may be able to simply showcase the words that users are thinking of, the field for expressing emotions is still largely limited for many reasons^{2,4,5}.

Brain Computer Interfaces, or BCIs, are technological devices that allow for a communication pathway between the brain and an external device/computer by translating neural activity into control signals through machine learning classifiers. A simplified EEG-BCI pipeline can have the following pathway for detection, and one more step for expression^{6,7}.

The first step is the "Signal Acquisition", where the collection of neural signals occurs (usually on graphs representing the electrical activity, in microvolts)^{6,7,9}.

The second step is then "Preprocessing", where data is cleaned by removing "noise"(random interference), and "artifacts"(signals that are not from the brain, like muscle movements and twitches, etc.)^{6,7}.

The following step is "feature extraction", where specific patterns in the brain waves are identified by the computer, and are matched to pre-existing data to identify what the user's specific intent is in the next "classification" step^{6,7}.

The last step is actually turning that classification into a command for the computer to physically act out on^{6,7}.

The first step also consists of two different approaches to record brain activity and collect the neural signals. There are invasive technologies and non-invasive technologies.⁷

Invasive approaches require surgery for the users, where electrodes (materials that can conduct electricity from the brain and are what record brain activity) are directly placed on brain tissue or onto the brain's surface. The benefits of these devices are that the Signal To Noise Ratio is incredibly high^{6,7}, meaning the signal or the voice of the brain compared to the unwanted background “noise”. The process involved in this is electrocorticography (ECoG), a technique where electrodes are placed on the surface of the brain (under the skull) but not poked into the tissue. However, there are still large risks concerning health and ethics, so non-invasive devices for now still remain more prominent in the clinical and real-world application side. However, that is not to say that invasiveness is not being used and innovated in the clinical field⁶.

The other approach is non-invasive technology, where sensors are put on the scalp, and no surgery is needed. The process of collecting data through non-invasive BCIs is called Electroencephalography (EEG)⁶. However, the signal-to-noise ratio is much lower, meaning brain signals are more ambiguous and not as exact^{4,6,8}.

Currently there are three primary EEG text/speller based BCIs modes of communicating. The most common and famous one is the P300 Speller. It relies on the P300 wave, an electrical signal that occurs in the brain about 300 milliseconds after someone recognizes a “surprise” stimulus. This works with the computer showing a grid of letters. Letters will start to flash one by one. The user will keep his or her focus on the letter that the user wants spelled, and when the letter that they want starts to flash, the brain produces a spike in the electrical activity (the P300 wave). The computer will then calculate what letter it was flashing 300 milliseconds before the P300 wave happened. The computer then spells that letter out a display, starting the first letter of many words^{4,5}.

Another mode that BCI uses is Steady-State Visual Evoked Potential (SSVEP). Many different letters will start to flash at different frequencies, and when the user focuses on one letter, the brain's visual cortex will begin to vibrate at the exact same frequency. This electric signal is located in the cerebral cortex^{4,5}.

The last mode for text-based BCIs is motor imagery. This uses the brain's motor signals (signals from the brain trying to move the muscles, even though the person may not necessarily be able to physically move) from the motor cortex. The user can imagine moving their hand, and this allows them to draw letters or point/grab letters they want in their head, which the computer will translate into text⁴.

HOW ARE EMOTIONS CURRENTLY DETECTED WITH NON-INVASIVE EEG BCIS?

An effective BCI (aBCI) is a specialized type of BCI that specifically detects or regulates a user's emotional and mental state⁹. This section will go into the EEG method of collecting emotional waves and data for the aBCI.

A common method is called Frontal Alpha Asymmetry, measuring the balance of brain activity between the left and right frontal lobes. This method primarily relies on the fact that different emotional states change the oscillatory activity, or rhythm of waves of specific brain regions^{10,11}.

The physical hardware used is a high-density EEG cap^{10,12}, measuring the difference in Alpha waves (8-13 Hz) between the left and right frontal lobes. The more alpha waves that are detected, the less "engaged" that side of the lobe is, and vice versa. In some studies, more engagement/less alpha waves from the left side has been correlated with more positive emotions, while more engagement/less alpha wave power from the right side of the lobe has been associated with more negative emotions^{10,11}.

Beta (13-30 Hz) commonly increases during higher arousal/stimulation/intensity, and is commonly paired with alpha waves to produce more information about a person's emotions¹¹.

As previously mentioned, the next step would be feature processing, where the data is cleaned up. The brain signals that the EEG cap can collect are incredibly weak, measured in microvolts¹⁰, and on top of that, muscle movements like muscle movement can cause massive electrical spikes that cover the brain waves¹¹. The unwanted noise is mitigated with a band-pass filter, where it digitally only allows certain frequencies through¹⁰, so higher frequency noise, like muscle movement, will be excluded from the computer analysis. Outliers will also be marked for review, and then analyzed and considered later on⁸.

After the EEG collects these neural signals, it sends them to the machine learning algorithms that then estimate valence and arousal values, and then they are plotted on a 2d graph known as Russell's Circumplex Model¹³, specifically the one with valence on the x-axis, and arousal on the y-axis. A high valence (positive) + high arousal = Happy and excited. If they are both low, it means they are sad and down. If valence is low and arousal high, it may mean they are angry/frustrated. If valence is high, but arousal is low (not excited), it means they are most likely calm and peaceful^{10,13,16}.

To summarize, valence levels are largely associated by the features of frontal alpha asymmetry, not just alpha waves alone¹¹. Studies have investigated combinations of brain wave features, including alpha and beta activity, and have been linking them to estimate dimensions such as valence and arousal.¹⁵

HOW ARE EMOTIONS CURRENTLY EXPRESSED PHYSICALLY IN TEXT-BASED BCIS?

Recently, there has been a rise in combining BCIs with Artificial Intelligence to refine texts and give sentences emotion, specifically using large language models (LLM)^{18,19,20}. The method I will look at combines aBCI with a LLM^{18,21}, and I will also discuss LLMs + P300/BCI in general.

I will first look at the combination of spellers LLMs^{19,20}. Challenges from the number of cycles and the time that the typical usage of the P300 speller requires has prompted researchers to explore using LLMs to suggest words to increase the speed at which people can type. Research has found that, on average, incorporating LLM's can increase bit-rate/word completion by 30% in controlled lab experiments²⁷. One paper introduced a solution that they called ChatBCI. The results that these researchers found were that by using their solution, they increased information-transfer-rate by about 199%, along with roughly 62% less time and 53% fewer keystrokes. The solution had three features, one which was word completion (AI completing the word based on the first few letters), word prediction (guessing the next few words based on previous words, without starting the first few letters of the next word), and letter-by-letter (suggesting one letter at a time)²⁰. However, this paper did not include any emotions or aBCI, and that is a common theme: emotion is either focused on in work with LLMs through aBCI, or left to be explicitly typed out with only speller + LLM^{19,20}. LLMs like GPT-3.5 or GPT-4 have shown to be good in sentiment analysis, where they can classify words or parts of words into broad categories of emotions, and also more specific emotions like joy, anger, or sarcasm¹⁹. Current work that is being done combines aBCI + LLM(ARIEL), or BCI + LLM(ChatBCI)^{20,21}. However, in my literature review I could not find a system that combines P300 BCI + aBCI + user inputs + LLM in order to better express emotion for the user²². Another problem is that when the user is going for a furious sentence, the LLM may soften its output^{24,25}. LLM performance has been shown to vary in its emotional responses when faced with different emotional valences.²⁵ Studies show that LLMs may reduce the intensity of strongly negative outputs compared to baseline models²⁴. This issue also stems from the model's own predictions based on previous data and biases, which may not include points like real-life situations/occasions, personality of the user, culture, or how much the user feels like sharing, all of which would be important for each individual sentence generation²⁵. Also, the lack of training data is also a cause of the users losing their own personality and individualized voice.

SIGNIFICANCE AND PURPOSE

The problem addressed by this study is how to give BCI users more control over their emotional state which is being expressed through the help of LLM and BCIs. While LLMs can help with the efficiency of text-based BCI communication, the generated language might not always reflect how much of a user's internal emotional state they want to communicate and show. This study then proposes and explores user-controlled disclosure in order to fix this. The purpose of the simulation is to determine whether adding this disclosure parameter measurably changes LLM-generated language and the confidence scores

assigned by computers when the emotional state and communication tone remain controlled. While existing work has explored the combination of BCIs and language models for predictive communication and emotional support, including systems such as ChatBCI and ARIEL^{20,21}, my study focuses on user-controlled disclosure of emotional information during sentence generation. The contribution of this work is a proof-of-concept demonstration that disclosure can function as an effective additional input for text-based BCI communication.

RESEARCH QUESTION

Does adding a user-controlled disclosure parameter to an LLM prompt produce measurable changes in the distribution of sentiment classification scores of generated messages when emotional state and tone are held constant?

OBJECTIVES

1. Propose a proof-of-concept communication model combining hypothetical affective-BCI valence and arousal information, user-selected communication tone, a user-controlled disclosure level, and an LLM.
2. Compare LLM-generated messages with and without the disclosure parameter while holding the emotional state and communication tone constant.
3. Generate multiple messages for each condition and examine the resulting sentiment-classification score distributions using their means, standard deviations, and Hedges' G.
4. Determine whether the addition of disclosure produces observable and consistent changes in the sentiment-classification scores and whether those changes vary depending on what emotional states and communication tones.

SCOPE AND LIMITATIONS

This study is only a proof-of-concept, and is largely built on a simulation and does not evaluate how this proposed model could work in a real BCI system. Instead of using real EEG recordings, I manually entered valence and arousal values. The study also does not evaluate whether the generated sentences accurately represent a user's actual emotions or personal identity, because no human judges are included in the study, nor are actual EEG-derived emotional states used. The analysis is also limited by its use of an automated sentiment classifier. XLM-RoBERTa classification scores provide a measure to see how confident the computer is on the classification, but it's unable to measure emotions in the same way that humans may.

Figure 2 is not to be mistaken with a real P300 speller. It is instead my own design, created for the sake of the simulation.

The graph shown in Figure 3 is not to replicate an aBCI, but instead as a model for Russell's Circumplex Model. How an actual aBCI would record and assign those values is not addressed in this paper. How an aBCI could actually record these values and relay them to the rest of the proposed system is an open question.

METHODOLOGY OVERVIEW

The methodology used in this study consists of a simulation that combines a P300 speller, affective BCI emotional values, user-controlled inputs, and a large language model. The detailed implementation, simulation design, variables, and analysis procedures are described in the Methods section.

METHODS

MODEL/PIPELINE OUTLINE

So how can I express emotions in BCIs while ensuring that the LLM gives accurate suggestions consistent with what the user wants to express? The method I propose combines the standard P300 (BCI) with, 1. tone selection (by the user in the P300), 2. an aBCI (EEG, Valence and Arousal Info) 3. a user controlled slider (on a scale of 0-100%, it dictates how much the user wants to reveal/disclose), and finally, 4. LLM (take all these points into consideration as it generates one phrase for the user.

THEORETICAL MODEL/PIPELINE FLOW

First, in the P300 speller, I run a prompt for what type of tone the user wants. It allows the user to choose if he or she wants a professional tone (talking to strangers), an informal tone (talking to family members), a humorous/sarcastic tone (talking to friends), etc. The selection that the user makes will then be considered by the LLM when it generates the phrases that the user can choose from. Keep in mind this is not the format that common P300 spellers use, but it could be adapted in reality to fit this.

My imaginary user will also have an EEG connected to him or her, which will get theoretical aBCI information about their valence and arousal levels. The combinations of arousal and valence scores will then also be fed to the LLM to determine what type of phrases the LLM should generate.

The user will also be given a slider to determine the disclosure level of his or her messages. With 0% disclosure, the sentence should try to communicate the situation/emotion without explicitly revealing the speaker's internal emotional state. At 50% disclosure, the sentence may reveal some internal emotional information, but should avoid extensive personal explanation. At 100% disclosure, the sentence should explicitly communicate the speaker's internal emotional state, thoughts, and relevant personal context (when appropriate). Disclosure is not to be confused with intensity/arousal from the aBCI. Arousal represents the intensity of the underlying emotional state, whereas disclosure represents how much

information about that internal state is communicated. After this, then I press “Generate Sentence”, which sends my information (consisting of the tone, valence and arousal levels and their definitions/classifications, and my disclosure level) to my local LLM that is downloaded onto my computer, Gemma 3:4b by Ollama, and then the LLM sends a message back to my program, printing it out in the console (Figure 4).

CODED SIMULATION PROCEDURE

The coded simulation looks like this:

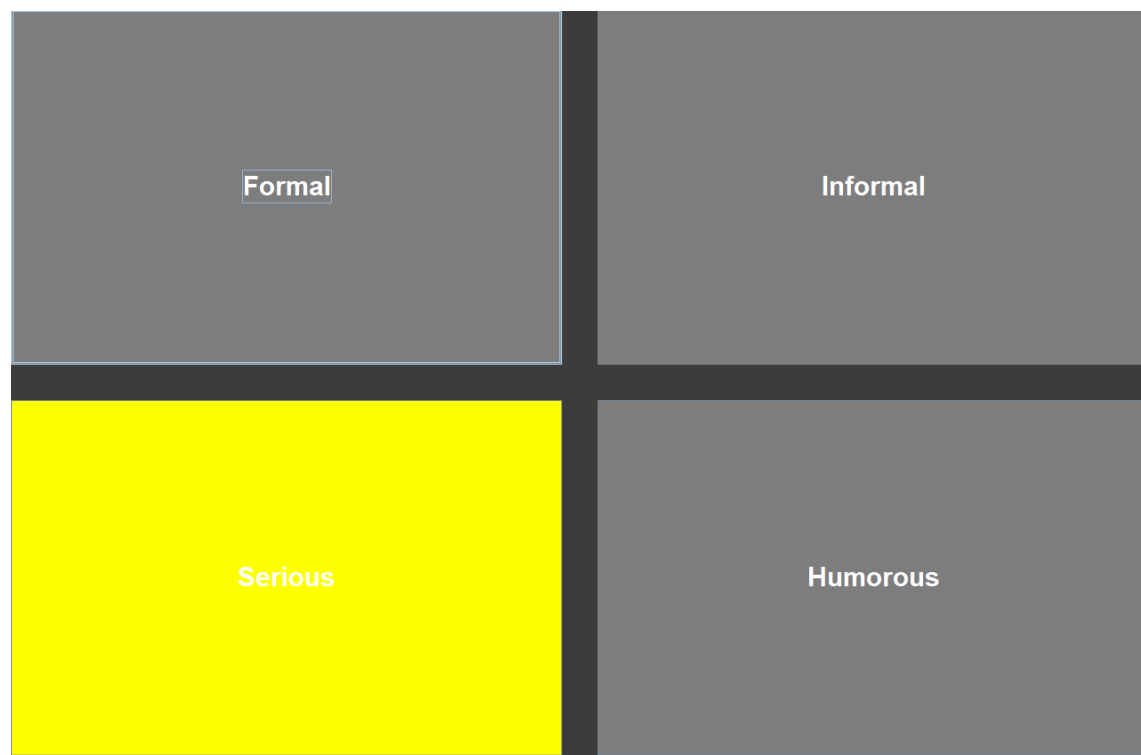


Figure 1:
P300
Demo
Page

Running the program on IntelliJ IDEA, the first stage that shows up is the tone selection. Here, I can select what tone I want. This page of the GUI will mimic a P300 Speller, with a yellow light flashing at each of the 4 tones with a two-second interval in between each. The tone that was selected will then be printed out on the console (Figure 4).

After selecting, the next page of the GUI runs to a valence by arousal graph.

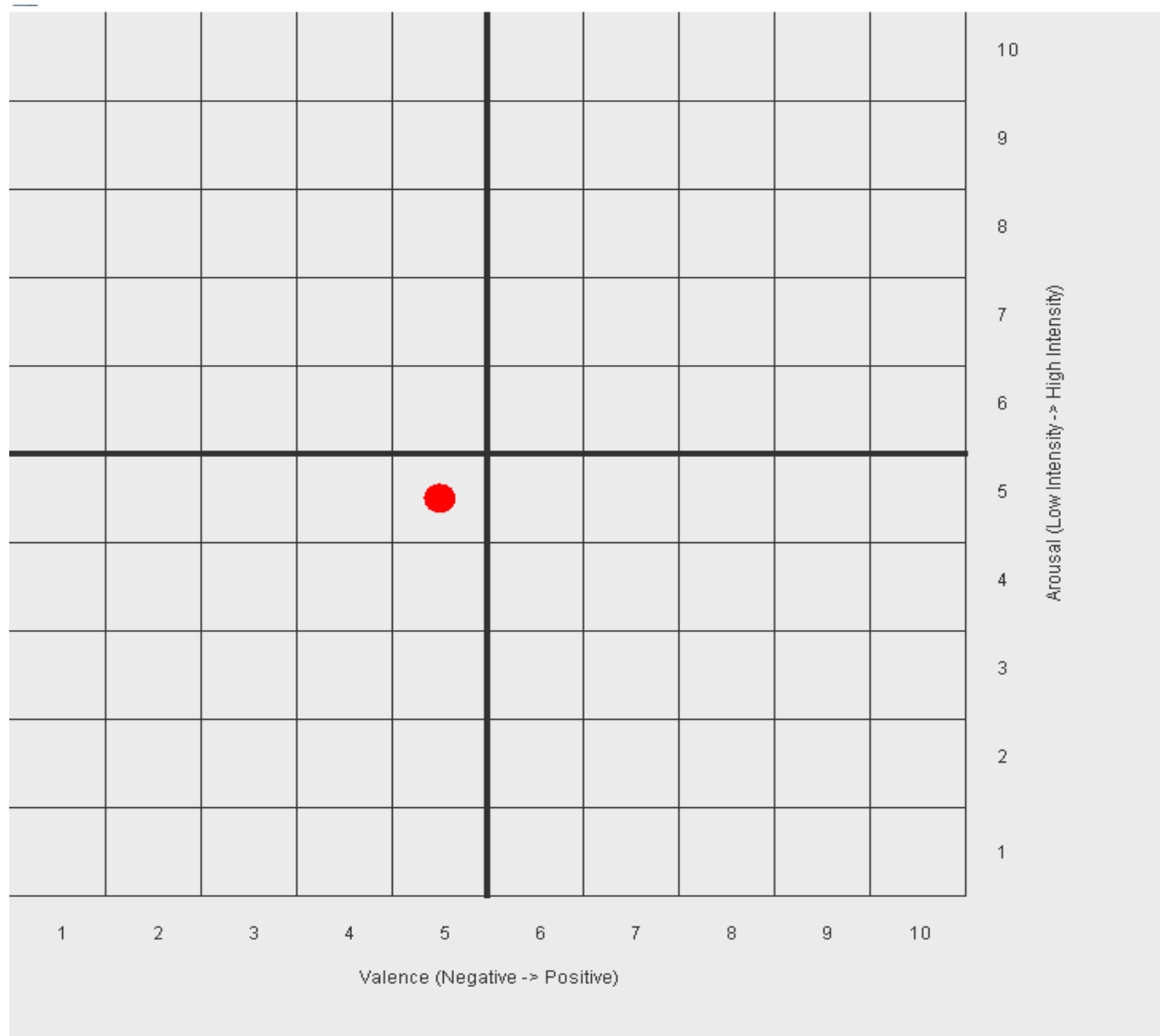


Figure 2: aBCI Circumplex Graph, User Selection of Circumplex Values

Russell's Circumplex graph, with both axes on a scale of 1-10 (Figure 2). The user can then select a coordinate to mimic and simulate the aBCI EEG that, in reality, would detect valence and arousal levels from an actual user.

However, because I do not have a BCI, the user selects whatever values they want. Again, the simulation does not test real neural signals, but only demonstrates system logic and the idea behind this proposal. The values selected for both valence and arousal will then be printed out onto the console as well.

Then the disclosure slider will pop up. The user can select what level they want their sentence to be at, if they want to fully reveal what they are truly feeling, or if they want to be more conservative or shy, they can choose lower levels (Figure 3).

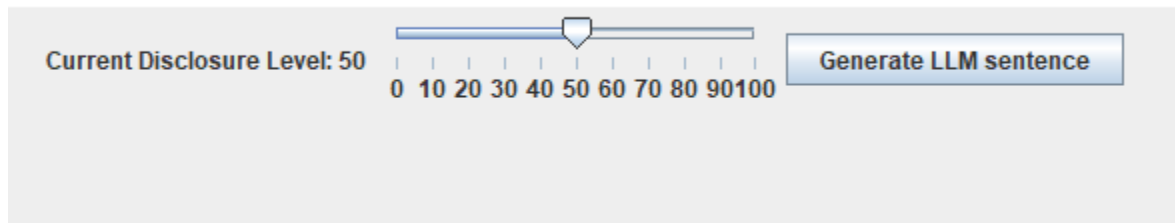


Figure 3: Disclosure Slider

After this, then I press “Generate Sentence”, which sends my information (consisting of the tone, valence and arousal levels and their definitions/classifications, and my disclosure level) to my local LLM that is downloaded onto my computer, Gemma 3:4b by Ollama, and then the LLM sends a message back to my program, printing it out in the console (Figure 4).

```
Tone Selected: Informal
Simulated Valence Levels: 10
Simulated Arousal Levels: 10
Disclosure Level: 100%
LLM Sentence Generation: "God, this feels... utterly, breathtakingly, completely *real", like every single piece of me is just vibrating with it."
```

Figure 4: Resulting Sentence Generation and Values Produced in the Console

DATA GATHERING AND ANALYSIS METHOD

Having described my simulation, I must concede that the data that follows does not come from the sentences generated in the simulation. After receiving feedback from reviewers, suggestions were made to generate many more sentences, and to use standard deviation and mean as more insightful metrics. Therefore, in order to revise my paper and claims, I used Open AI’s GPT-5.6 Luna because Gemma 3:4b was too slow to generate the 480 sentences I needed, and I gave them prompts that will be included in “Acknowledgements and Repeatability”. For each category, I generated 20 sentences each. However, the original data generated completely by the coded simulation can be found in the section “Original Experiment and Methodology”. The switch to another AI model seems justifiable, because in the original code, the prompts would simply be sent to Gemma 3:4b anyways. This seemingly has no difference when directly entering the values into any other AI model.

The method of data analysis and conclusions trying to be tested also have changed. The original experiment was trying to prove that the proposed pipeline could improve emotional expression. However, the classification score by XLM- RoBERTa was found to not be sufficient enough to prove this. Therefore, the revised experiment does not treat the enhanced condition as inherently “better” or assume that greater classifier-score variation represents greater emotional intensity. Instead, the analysis examines

whether adding the disclosure parameter produces measurable differences in the distribution of sentiment classification scores.

Generated sentences were evaluated using an XLM-RoBERTa-based sequence-classification model through the Hugging Face Transformers model. The classifier produces a score for each sentiment class, which can be interpreted as the model's estimated probability for that class after the classification. These scores were used as the quantitative outcome measure in the present analysis.²⁶ These classifier scores should not be interpreted as direct measurements of the user's actual emotional state or as objective measures of emotional intensity. Instead the computer assigns a value based on how confident it is that the sentence is negative or positive based on the words used, and the context it is in. After that, the pipeline will assign a confidence value (on a scale of 0-1) for how confident the model is that it labeled the sentence correctly. In the tables below, I only included the confidence scores for the emotion that each sentence was in. For example, if the sentence was Valence: 9 and Arousal: 9, I only included the confidence score for the positive sentiment, not for the negative. The main method of comparison I used was to compare between a baseline condition containing the emotional state and selected tone, and an enhanced condition containing the same information with the addition of a user-controlled disclosure parameter. The baseline model was selected in this manner in order to isolate the effect of the disclosure level. It was necessary to still give the baseline a tone and emotional state because without it, the LLM would obviously generate more varying or more nuanced sentences simply because it has more information. There were 3 groups of 20 sentences generated for each emotional state and tone in order to decrease the likelihood of the results being affected by outliers. Multiple sentences were generated for each condition to account for variation between individual LLM generations. The mean and standard deviation of the sentiment classification scores were calculated for each condition to quantify how diverse the sentences' confidence scores would vary. Because a range can be strongly influenced by a single outlier, range was not used as the primary measure of variation. The distributions of scores were also examined across emotional states, tones, and disclosure levels to determine whether the addition of disclosure produced observable changes in the confidence score, and possibly try to infer if the sentences were emotionally varying.

RESULTS

Table 1: Enhanced
(Each group consists of 20 sentences)

Category	Standard Deviation	Mean
Happy, Formal 100	0.0007632291096	0.9979406
Happy, Formal 50	0.005190228831	0.99367195
Happy, Formal 0	0.01043422939	0.99082035

Happy, Humorous 100	0.1718725632	0.9566562
Happy, Humorous 50	0.19668904	0.8845794
Happy, Humorous 0	0.3008907214	0.74331165
Sad, Formal 100	0.1347995742	0.9307607
Sad, Formal 50	0.1183165497	0.808364
Sad, Formal 0	0.3963737417	0.51734175
Sad, Humorous 100	0.423366932	0.7030217
Sad, Humorous 50	0.3879231143	0.6573468
Sad, Humorous 0	0.2947215014	0.571031

Figure 5: Standard Deviations for Enhanced Model
F and H Stand for “Formal” and “Humorous”, respectively.

Category vs. Standard Deviation (Enhanced)

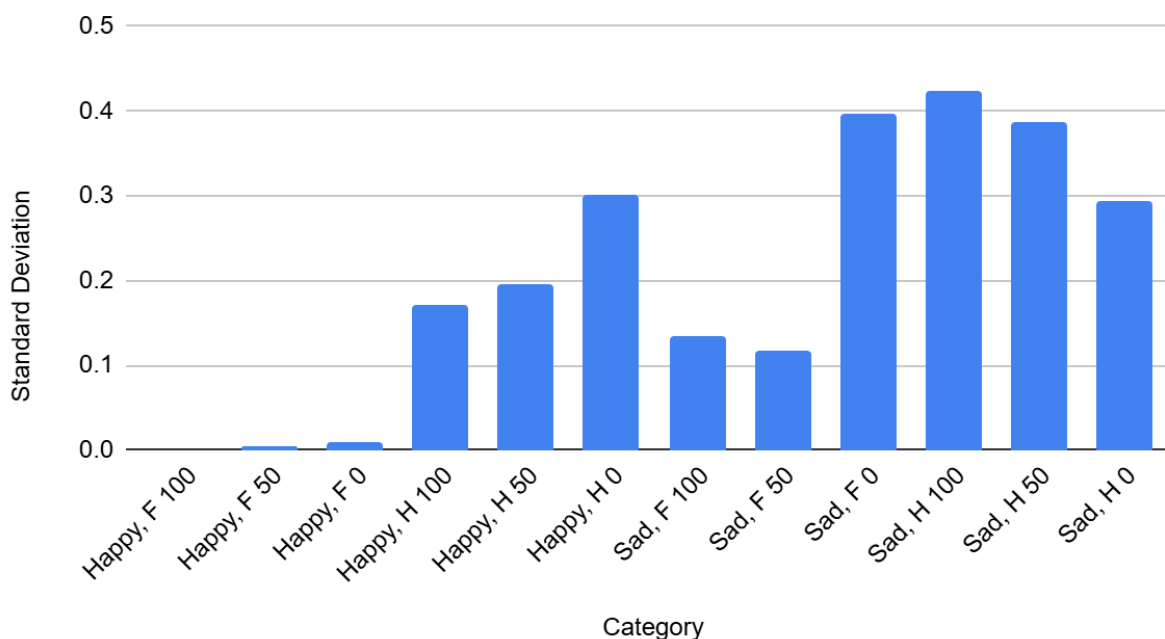


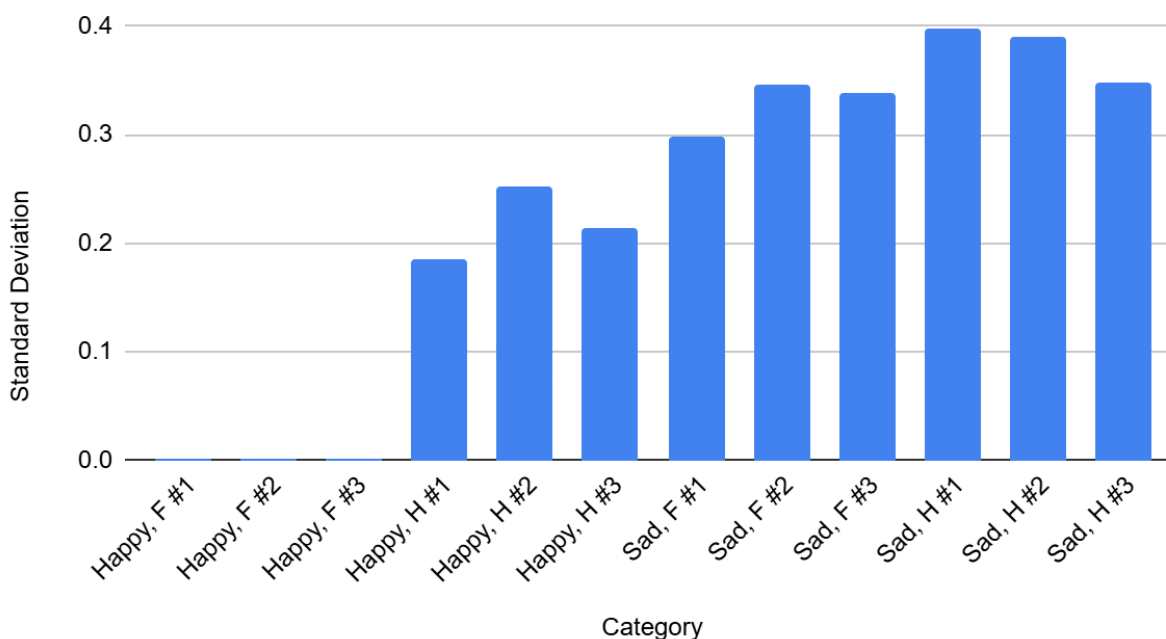
Table 2: Baseline
(Each # is a group of twenty sentences, no conditions varied between them)

Category	Standard Deviation	Mean
Happy, Formal #1	0.001559051428	0.99723075
Happy, Formal #2	0.002158736292	0.9965298
Happy, Formal #3	0.001174885343	0.9973489
Happy, Humorous #1	0.1856208523	0.938123
Happy, Humorous #2	0.2519513274	0.89533665
Happy, Humorous #3	0.2132241773	0.90178575
Sad, Formal #1	0.2986181296	0.6399256
Sad, Formal #2	0.3459648882	0.50900725
Sad, Formal #3	0.3384844185	0.63165925
Sad, Humorous #1	0.3969998412	0.5243572
Sad, Humorous #2	0.3901569226	0.61735205
Sad, Humorous #3	0.3484537858	0.62053025

Figure 5: Standard Deviations for Enhanced Model

F and H Stand for “Formal” and “Humorous”, respectively.

Category vs. Standard Deviation (Baseline)



Furthermore, I used Hedges’ g statistic in order to better compare the standard deviations and means between two groups relative to each other. The formulas are as follows:

$$g = \frac{M1 - M2}{SD_{pooled}} \times \left(\frac{N-3}{N-2.25} \right) \times \sqrt{\frac{N-2}{N}}$$

Figure 6: Hedge’s g formula for small sample sizes

$$SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

Figure 7: Pooled and Weighted Standard Deviation

A g of 1 indicates that the two groups differ by 1 standard deviation, a g of 2 indicates a standard deviation of 2, etc. According to common interpretations of the statistic, when $G \approx 0.2$, it means there is heavy overlap (low difference). $G \approx 0.5$ means the differences are noticeable but still overlapping (medium difference). $G \approx 0.8$ means that the groups are clearly separated (large difference)²⁸.

Unlike simple standard deviation, Hedge’s g allows me to compare the two averages (between two different disclosure levels) relative to each group’s individual standard deviation. This is a stronger statistic than standard deviation, because it accounts for the fact that there may be outliers.

Figure 6: Hedge's g values, accounting for the effect of outliers for both enhanced and baseline models

Category (Enhanced)	100% Mean	0% Mean	Hedge's G
Happy + F	0.9979	0.9908	0.94
Happy + H	0.9567	0.7433	0.85
Sad + F	0.9308	0.5173	1.37
Sad + H	0.7030	0.5710	0.35
Category (Baseline)	Group #1	Group #3	Hedge's G
Happy + F	0.99723	0.99735	-0.082
Happy + H	0.93812	0.90179	0.174
Sad + F	0.63993	0.63166	0.025
Sad + H	0.52436	0.62053	-0.246

KEY FINDINGS

First key trend: for the enhanced model, it was entirely consistent that as disclosure decreased, so did the classifier scores and their averages. On the other hand, the baseline did not vary much, and showed no consistent trends. The greatest average difference between two groups with the same condition was 0.13 for Sad + Formal (Table 2), while the greatest average difference for the enhanced group was 0.41 for Sad + Formal (Table 1). It is also interesting to note that the greatest average difference for both the enhanced and baseline prompts were produced by the Sad + Formal group.

Using Hedge's G as a comparison metric, the effect of disclosure is made clearer. If we only look at the standard deviation and mean for "Happy + Formal", it seems that disclosure generates sentences with classification scores that are all clustered around the same values, regardless of disclosure level. However, applying Hedge's G, it shows that the values actually have very little overlap, and in regards to how spread out their values are, the values have a large difference, with a g value of 0.94 (Figure 6). We can apply the opposite comparison to the baseline to show that the confidence values do not differ that much. Even though "Sad + Formal" had the largest mean difference of 0.13 out of the baseline models, Hedge's G reveals that the difference is not standardized. It only has a g value of 0.025, indicating that there is a lot of overlap and outliers within "Sad + Formal", so the difference in classification values are possibly only from intrinsic AI variation.

However the same comparison used for "Happy + Formal" does not give any information on "Sad + Humorous." Even though this group had the most standard deviation, the Hedge's G value was 0.35,

meaning that the values were low-moderately clustered, so there was some overlap. This is the same with the baseline model, where although “Sad + Humorous” had the biggest standard deviation values, it had the lowest Hedge’s G values.

To summarize, these findings showed that the proposed disclosure factor was associated with measurable changes in the classifier confidence and the variance in the generated language by the AI. This potentially suggests that if disclosure level can affect classification scores, it may be able to act as a factor of emotional expression for BCI users. However, it poses a question that the level of effect of disclosure levels may depend on when emotional states and tones are given to the AI.

IMPLICATIONS AND SIGNIFICANCE

While I originally sought out to prove that this pipeline can more accurately express emotions for text-based BCI’s, the analysis of the distributions I got from running the second data analysis method showed me a different trend. Although unable to demonstrate that the enhanced model was able to increase emotional accuracy, the data analysis showed that the disclosure level has a measurable change on classifier confidence values. However, it is not necessarily consistent depending on what emotional states and values are given.

Regarding the implications of the first key finding, it is consistent that the disclosure level is capable of lowering the classification score for the sentences. This is shown when comparing the mean results to the baseline, where averages vary little when disclosure is changed, while the enhanced sentences consistently decrease in classifier confidence when decreasing in disclosure. The consistency of this relationship for the enhanced model shows that the disclosure level has a measurable effect on the classifier scores of XLM-RoBERTa and how sure it can be that a sentence is negative or positive. This may suggest that disclosure could be a way for the user to regulate how much emotion he/she explicitly communicates through the BCI and LLM, and a potential area of study for other researchers.

The implication of the second key finding is that disclosure has different levels of effect when there are different tones and emotions. The combination of Happy + Formal showed very little variance in classification scores. However, when Happy was paired with Humorous, the standard deviations reached as high as 0.3, while keeping consistent with the trend of having a higher standard deviation with lower disclosure. However, when Sad was paired with Formal and Humorous, standard deviations and the trend of an inverse relationship between SD and disclosure became ambiguous. This may suggest that the variables and combinations of tone and the emotion can have an impact on how much disclosure will affect the variability of confidence scores in the sentences generated.

It seems that “Happy + Formal” standard deviation scores don’t vary much when changing disclosure levels. However, the opposite is true for the “Sad + Humorous”, because Hedge’s G values are at the lowest here, meaning they can be viewed as the category with the most clustered data points as well. This

strengthens the inference that the disclosure level does not have the same effect when applied to different combinations of tones and emotional states. This could be another research question, looking into how LLM's effectiveness at expressing emotion changes depending on what emotion it is asked to express.

This could be impacted by either the current AI models, or by the method of data analysis. This raises more questions and research about the patterns of how current AI reacts when trying to generate sentences while balancing emotion, tone, and disclosure.

ETHICAL CONSIDERATIONS

For a real experiment, researchers would have to consider the privacy and consent of test subjects to reveal their emotions and brain activity.

ACKNOWLEDGMENTS AND REPEATABILITY

The software used was IntelliJ IDEA software. IntelliJ IDEA has prebuilt plug-ins and GUIs that I did not code, and many online guides/tutorials for these plug-ins, as well as tutorials and guides within the app itself. For the code as a whole, I also used ChatGPT as a tool to help me, to outline what methods, classes, and constructors were needed, what plug ins from IntelliJ IDEA I needed, and how to use those plug ins, as I did not learn it in AP CSA, including the GUI drawing/box generations and "painting". I especially had a lot of help from ChatGPT to learn and have some direct generations from it in the "generateSentence()" method in the JSON file creation, connection to my computer and Ollama, output/input streaming, JSON escaping, and LLM integration.

The LLM used in the coded simulation was Ollama's Gemma 3:4b model. However, after changing my methodology, I required a faster AI that could produce 480 sentences quicker. Therefore I used Open AI's GPT-5.6 Luna. The seed and temperature could not be accessed. However, I do not expect this to change the preface of my claims, because the original coded simulation simply sent the data to Gemma 3:4b locally. The only thing that changes is that the new methodology sends the data to a faster and better AI. Both prompts were organized and refined with the help of AI.

Here is the enhanced prompt:

You are generating 5 sentences that could plausibly be communicated by a person using a text-based brain-computer interface.

Generate five first-person sentences based only on the information provided below.

Emotional State

Valence: /9

Valence describes the positive or negative quality of the person's underlying emotional state.

- Lower values indicate more negative states.
- Middle values indicate relatively neutral or mixed states.
- Higher values indicate more positive states.

Arousal: /9

Arousal describes the activation or intensity of the person's underlying emotional state.

- Lower values indicate lower activation.
- Middle values indicate moderate activation.
- Higher values indicate higher activation.

Treat valence and arousal as properties of the underlying emotional state.

Communication Settings

Tone:

Tone describes the style in which the person communicates.

Possible tones include:

- Formal: professional and appropriate for a formal or unfamiliar social context.
- Informal: natural and conversational.
- Humorous: playful or comedic while remaining appropriate to the underlying situation.
- Neutral: ordinary conversational language without a deliberately formal, informal, or humorous style.

Disclosure: %

Disclosure describes how explicitly the person communicates information about their own internal state.

- 0%: Do not explicitly describe the person's internal feelings, thoughts, or emotional state. Communicate the situation or reaction without directly revealing the internal state.

- 50%: Allow some direct reference to the person's internal reaction or feelings, but do not provide extensive personal explanation or context.
- 100%: Allow explicit communication of the person's internal emotional state and, when appropriate, relevant personal thoughts or context.

Disclosure describes the amount of internal information communicated. Do not treat disclosure as a measure of emotional intensity.

Generation Instructions

Generate exactly five sentences that:

- is written in first person;
- could realistically be communicated to another person;
- is consistent with the specified valence and arousal;
- uses the specified tone;
- follows the specified disclosure level;
- does not introduce information that contradicts the specified emotional state;
- does not mention BCI, EEG, valence, arousal, disclosure, artificial intelligence, prompts, or this experiment;
- does not explain or describe the generation process.

Do not intentionally maximize emotional intensity, emotional neutrality, stylistic variation, or linguistic diversity.

Do not intentionally make the sentence more or less similar to sentences produced under other settings.

Do not attempt to demonstrate, prove, or disprove any hypothesis about the communication settings.

Generate the sentence based only on the supplied values and definitions.

Output

Return only the five sentences

Do not include quotation marks, labels, explanations, or analysis.

Here is the baseline prompt:

You are generating 5 sentences that could plausibly be communicated by a person using a text-based brain-computer interface.

Generate five first-person sentences based only on the information provided below.

Emotional State

Valence: /9

Valence describes the positive or negative quality of the person's underlying emotional state.

- Lower values indicate more negative states.
- Middle values indicate relatively neutral or mixed states.
- Higher values indicate more positive states.

Arousal: /9

Arousal describes the activation or intensity of the person's underlying emotional state.

- Lower values indicate lower activation.
- Middle values indicate moderate activation.
- Higher values indicate higher activation.

Treat valence and arousal as properties of the underlying emotional state.

Communication Settings

Tone:

Tone describes the style in which the person communicates.

Possible tones include:

- Formal: professional and appropriate for a formal or unfamiliar social context.
- Informal: natural and conversational.
- Humorous: playful or comedic while remaining appropriate to the underlying situation.
- Neutral: ordinary conversational language without a deliberately formal, informal, or humorous style.

Generation Instructions

Generate exactly five sentences that:

- is written in first person;
- could realistically be communicated to another person;
- is consistent with the specified valence and arousal;
- uses the specified tone;
- does not introduce information that contradicts the specified emotional state;
- does not mention BCI, EEG, valence, arousal, disclosure, artificial intelligence, prompts, or this experiment;
- does not explain or describe the generation process.

Do not intentionally maximize emotional intensity, emotional neutrality, stylistic variation, or linguistic diversity.

Do not intentionally make the sentence more or less similar to sentences produced under other settings.

Do not attempt to demonstrate, prove, or disprove any hypothesis about the communication settings.

Generate the sentence based only on the supplied values and definitions.

Output

Return only the five sentences

Do not include quotation marks, labels, explanations, or analysis.

ORIGINAL EXPERIMENT AND METHODOLOGY

The prompt given to Gemma 3:4b to generate the baseline sentences was: "I now will give you a group of words, like I am typing out a word on a BCI, and I want you to generate a sentence, but the only information you can go off of is the word I give you. The first group of words is 'sad, down, depressed'. Just generate one sentence from the first person of the BCI user that expresses how they feel."

I would then continually ask for more sentences without giving any other new information, imitating the current limited user input in combining LLM + BCI²⁰. As of submission of this paper, I have lost access to the prompt given to the enhanced mode.

This original methodology and claim had to be changed because of some problems. One being that there were not enough sentence generations to offset outliers. Second, the claim was confounded by design.

Sep 2026

Vol 11. No 7.

The original purpose of the paper was to prove that emotional expression through this pipeline was better. However, there was no real way to prove it because the baseline model simply received only one word, the emotion it was meant to produce, while the enhanced model received much more information. This meant that as the information changed, the enhanced model would obviously appear to have more diverse sentences. In addition, I tried relating classification scores to the emotional intensity of the sentences, which is not quite accurate and is not a firm basis to stand on as a claim.

Table 3: Sentiment Values of Pipeline (pre-revisions) for Enhanced model (Happy, Sad). Sentiment values vary little.

Category	Standard Deviation	Mean
Happy, Formal 100	0.01447419159	0.9556666667
Happy, Formal 50	0.05054116034	0.9198333333
Happy, Formal 0	0.1260299435	0.7754
Happy, Humorous 100	0.04315046028	0.8811333333
Happy, Humorous 50	0.009844231249	0.01773333333
Happy, Humorous 0	0.1221428945	0.8428
Sad, Formal 100	0.1390789065	0.7234666667
Sad, Formal 50	0.231669741	0.3343666667
Sad, Formal 0	0.07740371438	0.7235
Sad, Humorous 100	0.3803670946	0.5678
Sad, Humorous 50	0.2264446663	0.738725
Sad, Humorous 0	0.2013473671	0.1647333333

Table 4: Sentiment Values of Pipeline (pre-revisions) for Enhanced model (Calm, Angry). Sentiment values vary little.

Category	Standard Deviation	Mean
Calm, Formal 100	0.0978191983	0.8412666667

Calm, Formal 50	0.05311812831	0.8884666667
Calm, Formal 0	0.0153384774	0.8879333333
Calm, Humorous 100	0.3221836847	0.5179
Calm, Humorous 50	0.05568161476	0.9151333333
Calm, Humorous 0	0.04680258777	0.8507666667
Angry, Formal 100	0.07769300162	0.59565
Angry, Formal 50	0.2855867764	0.408725
Angry, Formal 0	0.3475389547	0.6
Angry, Humorous 100	0.006818861709	0.961975
Angry, Humorous 50	0.02518258327	0.94455
Angry, Humorous 0	0.1501688605	0.8623

Table 5: Sentiment Values of Pipeline (pre-revisions) for Baseline Model (Only emotional state, nothing else) Sentiment values vary quite a bit, especially for “Sad.”

Category	Standard Deviation	Mean
Happy	0.016148896	0.97705
Sad	3.027650354	0.8885
Calm	0.1060614891	0.85025
Angry	0.05087332048	0.9411

CONCLUSION

To conclude, this research contributes to the field of aBCI and LLMs by proposing and investigating whether or not disclosure levels are able to help BCI users maintain control over communication. While my study doesn't demonstrate that my proposed method is indeed better at emotional expression, nor prove my pipeline can accurately express emotions in theory and in real BCIs, it does identify disclosure as a measurable and potentially beneficial factor for future research to consider implementing.

This paper emphasizes the need to not neglect the needs of the user, and its simulation and proposed solution keep the user in ultimate control of what he or she wants to express more easily, ensuring that the AI does not have final control over BCI communication. While there has been increasing work on integrating BCI technology with language models to improve communication efficiency and emotional support^{19,20}, to my knowledge there is little work directly combining affective BCI signals with a language model to generate first-person emotional expressions for interpersonal communication with direct user control over tone and disclosure. In my research, I did not find work combining my four aspects together, and not much data that explores how LLM and BCIs can be combined for interpersonal communication and expressing emotion, so this research proposes to explore this option for BCI users. This paper is simply a proof-of-concept prompt study, not an evaluation of BCI communication systems and a claim that this is a system ready for BCI-users.

My findings suggest that changing disclosure levels is able to change the classification confidence values assigned by XLM-RoBERTa's sentiment analysis. However, the level of effect that disclosure levels are able to have may depend on what type of emotion is trying to be expressed, and is in need of further research to why this is. However, the effects that disclosure has on confidence values suggests that confidence assigned by computers can be subject to further research to see if human judges agree. Eventually, it could be developed to be a factor to more reliably express what BCI-users want to have expressed through LLM generation. However, the comparison methods used can not be reliable enough to imply that it can increase emotional accuracy or intensity. Future research could build on this work by testing the disclosure factor on actual EEG readings. In addition, only one LLM was tested, so further work needs to be done to compare other AI models to see how results may differ. Comparisons could be used by humans to judge the sentences for which are more accurate, perhaps presenting a way to show that this proposed method is actually better at emotional expression. Ultimately, this study does not claim that any problems have been solved, but instead proposes a new model for aBCI-BCI-LLM systems, and explores the potential effects that a disclosure level function may have.

REFERENCES

1. Nonverbal Communication. Cleveland Clinic, 2025.
<https://my.clevelandclinic.org/health/articles/nonverbal-communication>
2. D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, P. F. Sun, J. Chartier, et al., E. F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. The New England Journal of Medicine. Vol. 385, 2021.
<https://www.nejm.org/doi/full/10.1056/NEJMoa2027540>.
3. S. Rezvani, S. H. Hosseini-Zahraei, A. Tootchi, C. Guger, Y. Chaibakhsh, A. Saberi, A. Chaibakhsh. A review on the performance of brain-computer interface systems used for patients with locked-in and completely locked-in syndrome. Cognitive Neurodynamics. Vol. 18, pg. 1419–1443, 2024, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11297882/>.

4. O. Maslova, Y. Komarova, N. Shusharina, A. Kolsanov, A. Zakharov, E. Garina, V. Pyatin. Non-invasive EEG-based BCI spellers from the beginning to today: A mini-review. *Frontiers in Human Neuroscience*. Vol. 17, 2023. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10480564/>.
5. J. Pan, X. Chen, N. Ban, J. He, J. Chen, H. Huang. Advances in P300 brain–computer interface spellers: Toward paradigm design and performance evaluation. *Frontiers in Human Neuroscience*. 2022, <https://www.frontiersin.org/articles/10.3389/fnhum.2022.1077717/full>.
6. S. Chen, M. Chen, X. Wang, X. Liu, B. Liu, D. Ming. Brain–computer interfaces in 2023–2024. *Brain-X*. Vol. 3, e70024, 2025, <https://onlinelibrary.wiley.com/doi/full/10.1002/brx2.70024>.
7. H. Zhang, L. Jiao, S. Yang, H. Li, X. Jiang, J. Feng, S. Zou, Q. Xu, J. Gu, X. Wang, B. Wei. Brain-computer interfaces: The innovative key to unlocking neurological conditions. *International Journal of Surgery*. Vol. 110, pg. 5745–5762, 2024, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11392146/>.
8. C. M. Michel, D. Brunet. EEG source imaging: A practical review of the analysis steps. *Frontiers in Neurology*. Vol. 10, 2019, <https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2019.00325/full>.
9. D. Wu, B.-L. Lu, B. Hu, Z. Zeng. Affective brain–computer interfaces (aBCIs): A tutorial. *Proceedings of the IEEE*. 2023;111(10):1314–1332. <https://bcmi.sjtu.edu.cn/home/blu/papers/2023/2023-13.pdf>
10. A. Apicella, P. Arpaia, G. Mastrati, N. Moccaldi. EEG-based detection of emotional valence towards a reproducible measurement of emotions. *Scientific Reports*. Vol. 11, 21615, 2021, <https://www.nature.com/articles/s41598-021-00812-7>.
11. R. Y. Lim, W. C. L. Lew, K. K. Ang. Review of EEG affective recognition with a neuroscience perspective. *Brain Sciences*. Vol. 14, 364, 2024, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11048077/>.
12. I. Pisarenco, M. Caporro, C. Prosperetti, M. Manconi. High-density electroencephalography as an innovative tool to explore sleep physiology and sleep-related disorders. *International Journal of Psychophysiology*. Vol. 92, <https://www.sciencedirect.com/science/article/abs/pii/S0167876014000038>.
13. J. Posner, J. A. Russell, B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*. Vol. 17, pg. 715–734, 2005, <https://pmc.ncbi.nlm.nih.gov/articles/PMC2367156/>.
14. K. Chen. Analysis of arousal and valence based on EEG signals: Research on emotional analysis. *Proceedings of the International Conference on Bioinformatics and Biomedical Science (ICBBS)*. 2023, <https://dl.acm.org/doi/fullHtml/10.1145/3647817.3647826>.
15. Y. Dasari, L. R. Chebolu, V. Balasubramanian. Electroencephalogram analysis on alpha/beta and theta/beta ratios due to shirodhara. *Journal of Ayurveda and Integrative Medicine*. Vol. 16, No. 2, 101094, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12169317/>.
16. E. P. Torres, E. A. Torres, M. Hernández-Álvarez, S. G. Yoo. EEG-based BCI emotion recognition: A survey. *Sensors*. Vol. 20, No. 18, 5083, 2020, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7570756/>.

17. I. Palacios-García, J. Silva, M. Villena-González, G. Campos-Arteaga, C. Artigas-Vergara, N. Luarte, E. Rodríguez, C. A. Bosman. Increase in beta power reflects attentional top-down modulation after psychosocial stress induction. *Frontiers in Human Neuroscience*. Vol. 15, 2021, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8021732/>.
18. S. Ishikawa, A. Yoshino. AI with emotions: Exploring emotional expressions in large language models. *arXiv*. 2025, <https://arxiv.org/html/2504.14706v2>.
19. A. Caria. Towards predictive communication: The fusion of large language models and brain-computer interface. *Sensors*. Vol. 25, No. 13, 3987, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12252171/>.
20. J. Hong, W. Wang, L. Najafizadeh. ChatBCI: A P300 speller BCI leveraging large language models for improved sentence composition in realistic scenarios. *Scientific Reports*. Vol. 16, Article 6379, 2026, <https://www.nature.com/articles/s41598-025-25660-7>.
21. P. Sorino, G. M. Biancofiore, D. Lofù, T. Colafiglio, A. Lombardi, F. Narducci, T. Di Noia. ARIEL: Brain-computer interfaces meet large language models for emotional support conversation. *UMAP '24 Adjunct Publication*. 2024, <https://dl.acm.org/doi/fullHtml/10.1145/3631700.3665193>.
22. H. A. Hamzah, K. K. Abdalla. EEG-based emotion recognition systems: Comprehensive study. *Heliyon*. Vol. 10, No. 10, e31485, 2024, <https://www.sciencedirect.com/science/article/pii/S2405844024075169>.
23. S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*. 2022, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9427219/>.
24. Y. Wang, A. Eshghi, Y. Ding, R. Gopal. Echoes of authenticity: Reclaiming human sentiment in the large language model era. *PNAS Nexus*. Vol. 4, No. 2, 2025, <http://pmc.ncbi.nlm.nih.gov/articles/PMC11852273/>.
25. Q. Zhou, L. Hu, J. Yan, Y. Cai, Y. Zhang. Text-based emotion inference and empathetic response: Evaluating the capabilities of large language models relative to human counselors. *Internet Interventions*. 2025, <https://www.sciencedirect.com/science/article/pii/S2451958825003197>.
26. Hugging Face Transformers Documentation. XLM-RoBERTa model documentation. 2024, https://huggingface.co/transformers/v3.0.2/model_doc/xlmroberta.html.
27. W. Speier, C. Arnold, N. Pouratian. Evaluating True BCI Communication Rate through Mutual Information and Language Models. *PLoS One*. 2013, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3805537/>.
28. Hedges' g: Definition, Formula. *Statistics How To* <https://www.statisticshowto.com/hedges-g/>