



## OPEN EEG reveals the cognitive impact of polarized content in short video scenarios

Bangde Du<sup>1</sup>, Ziyi Ye<sup>1</sup>, Monika Jankowska<sup>2</sup>, Zhijing Wu<sup>3</sup>, Qingyao Ai<sup>1</sup>✉, Yujia Zhou<sup>1</sup> & Yiqun Liu<sup>1</sup>✉

Short video browsing is a dominant medium today, significantly influencing public opinion and social dynamics. We investigate the effects of video platform Content Polarization (CP) on user perceptions and behaviors by bringing together measurements such as responses to self-report questionnaires, behavior signals, and noninvasive Electroencephalogram (EEG) signals. Our study demonstrated that just one week of exposure to polarized content can result in distinct patterns of behavior and brain activity during the perception of videos. Further, we revealed that exposure to content with diverging polarization directions, i.e. positive or negative, has extremely different effects on participants' emotional states, sentiment judgments, and brain responses. Specifically, participants' sentiment scores showed a significant decrease ( $Mean_{diff} = -0.385$ ,  $p = 2.35e^{-5}$ ,  $r = 0.484$ ) after exposure to negative content, whereas positive content led to an average sentiment score increase ( $Mean_{diff} = 0.171$ ,  $p = 0.046$ ,  $r = 0.308$ ). Finally, we show that analyzing participants' brain responses is even more effective at detecting exposure to polarized content than self-report judgments or behavioral signals, achieving superior predictive accuracy ( $ACC = 0.641$ ,  $F1 = 0.639$ ,  $AUC = 0.672$ ).

**Keywords** Content Polarization, Short Video Browsing, EEG

While online social media platforms have become key spaces for the distribution of information and the formation of public opinions, they have also been criticized for contributing to the polarization of opinions in politics<sup>1,2</sup>, healthcare<sup>3</sup>, and science<sup>4</sup>. One significant cause of this phenomenon is the content recommendation models employed by social media platforms, which encourage the users to interact with the contents based on their previous interactions and social connections<sup>5,6</sup>. The prevalence of such recommendation mechanisms has provoked concerns about users' limited exposure to opinions diverging from contents they have previously engaged with<sup>7,8</sup>. This dynamic has been described with terms such as "opinion polarization"<sup>5</sup>, "echo chambers"<sup>9</sup>, "filter bubbles"<sup>8</sup>, "selective exposure"<sup>10</sup>, and "information cocoons"<sup>11</sup> and linked to a series of social issues such as political polarization.

Although the above concepts have been widely used to critique modern online platforms, it is still difficult to quantify and conduct quantitative research on this phenomenon due to the lack of unified definitions<sup>12,13</sup>. Previous studies have attempted to measure polarization on social media by analyzing users' clicks, likes, and other behavioral data contained in social media platforms' logs<sup>14–18</sup>. Other research proposed strategies to alleviate polarization through diversification of the topics and views people encounter online through such methods as more exploratory recommendation systems<sup>19,20</sup>, presenting the users with opposing views<sup>21,22</sup>, and encouraging group discussions<sup>23</sup>. However, the above study assumes that users' online behavior can accurately reflect the degree of polarization. Unfortunately, research like those by<sup>24</sup> demonstrates that users' clicks and likes in response to polarized content do not necessarily reflect their opinions or attitudes towards such content. In contrast, other studies gathered users' explicit responses to polarized content relied on methods such as interviewing<sup>25</sup>, surveys<sup>26</sup>, and analyzing content posted by individual users<sup>27</sup>. Their research has demonstrated that polarization can affect an individual's sentiment<sup>24,25</sup> and emotional state<sup>26,27</sup>. Nevertheless, while the survey results show the overall trends of people's opinion changes, they can hardly provide detailed insights on how Content Polarization (CP) affects humans' subconsciousness and neural cognition.

Recent studies have shown that the brain signals of users when interacting with an information system can serve as important measurements for multiple information accessing factors, such as relevance judgment<sup>28,29</sup>,

<sup>1</sup>Department of Computer Science and Technology, Tsinghua University, Zhongguancun Laboratory, Haidian District, 100084 Beijing, China. <sup>2</sup>Department of Anthropology, Rice University, 6100 Main St, 77005 Houston, TX, USA. <sup>3</sup>Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, 100081 Beijing, China. ✉email: aiqingyao@gmail.com; yiqunliu@tsinghua.edu.cn

information needs<sup>30</sup>, and knowledge state<sup>31</sup>. As a portable and economical device for collecting brain signals, electroencephalogram (EEG) has been a widely used neurological method for measuring and understanding users' perception, attention, memory, and emotional state during information processing<sup>32</sup>. EEG signals can be collected during the task process without interfering with the user. Additionally, EEG can detect subtle changes in brain activity that may not be detectable through traditional behavioral metrics, providing a deeper insight into the cognitive and emotional processes underlying the process of opinion polarization. We believe EEG could complement or even serve as an alternative to other behavioral signals in opinion analysis, and provide important insights for understanding the impact of CP people's opinions.

In this paper, we propose to measure the influence of online content polarization, particularly those on short video browsing platforms, on users' opinion formation using both behavioral data and EEG measurements. We detect the sentiment and emotion shifts induced by exposure to specific content, showing that online content can drive people's opinions toward greater polarization. We aim to understand how content polarization in short video browsing influences users' sentiment judgments and brain signal patterns, and address the following questions: 1) How does Content Polarization (CP) in short video browsing affect users' sentiment judgment? 2) How does content polarization exposure affect users' brain signal patterns? 3) Can we predict the possibility of users' exposure to polarized short videos based on user behaviors and brain signals?

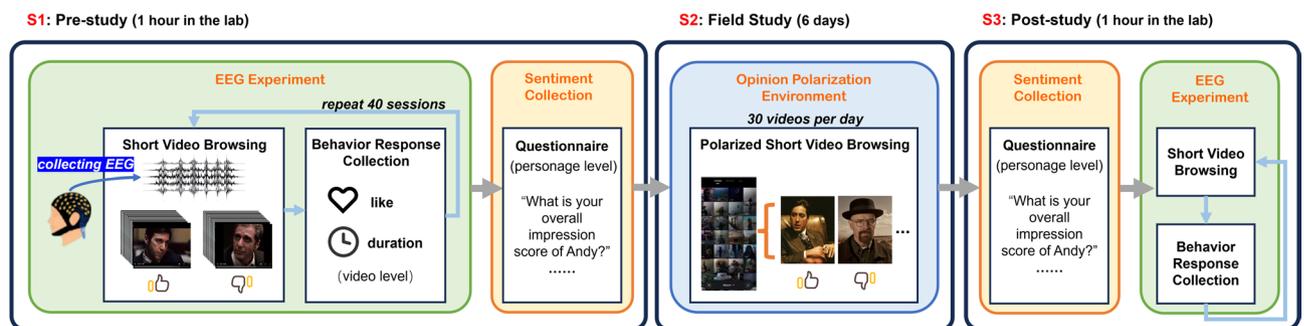
To address these questions, we conducted a three-stage user study, as shown in Fig. 1. A web-based short video browsing platform was built (see Fig. 2) to mimic modern social media and recommendation systems. The platform collected users' explicit sentiment judgments, behavior logs, and brain responses while watching short video stimuli before and after exposure to polarized content. Based on the collected data, we first explored how participants respond differently to short video content after exposure to polarized content with either positive or negative opinions. The experimental data showed that polarized contents with positive or negative opinions have a differential impact on participants according to their self-reports and behavior responses. Especially, we found that exposure to either positive or negative content sways participants' opinions towards the corresponding polarity. However, negative content exerts a more pronounced effect in terms of participants' sentiment judgment. In contrast, exposure to positive content elicits a greater impact on users' emotional state, but excessive positive content can lead to information fatigue and a rebound effect in terms of arousal. We further explored the connection between brain activity and both emotions and behavior in the context of short video browsing. We observed that CP is correlated with EEG, emotions, and behaviors to varying degrees. Based on this observation, we further utilized these data to predict the polarity (positive or negative) of CP. We found that using EEG data as inputs achieves better predictive accuracy even compared to participants' explicit sentiment judgments.

## Results

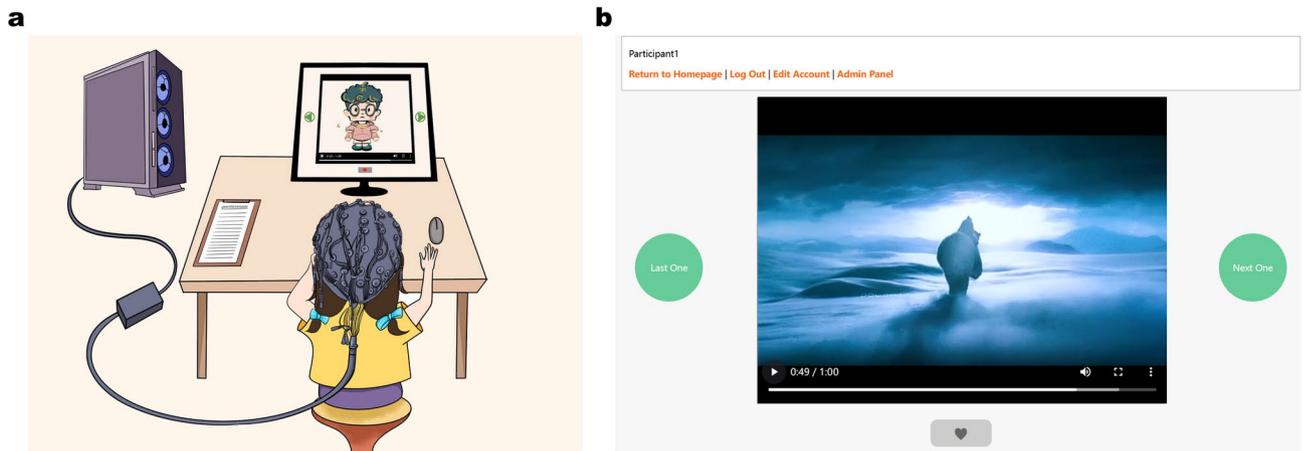
To assess the impact of polarized short video content, we recruited 24 participants, comprising 11 women and 13 men. We collected data on EEG signals, sentiment scores, behavioral responses such as like rates and view ratios, and emotional states, specifically valence and arousal, during both the pre-study and post-study phases.

As shown in Fig. 3, participants' sentiment scores showed a significant decrease ( $Mean_{diff} = -0.385$ ,  $w = 817.0$ ,  $p = 2.35e^{-5}$ ,  $r = 0.484$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ) after exposure to negative content (i.e., content with negative opinions on personage), whereas positive content (i.e., content with positive opinions on personage) led to an average sentiment score increase ( $Mean_{diff} = 0.171$ ,  $w = 302.5$ ,  $p = 0.046$ ,  $r = 0.308$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ).

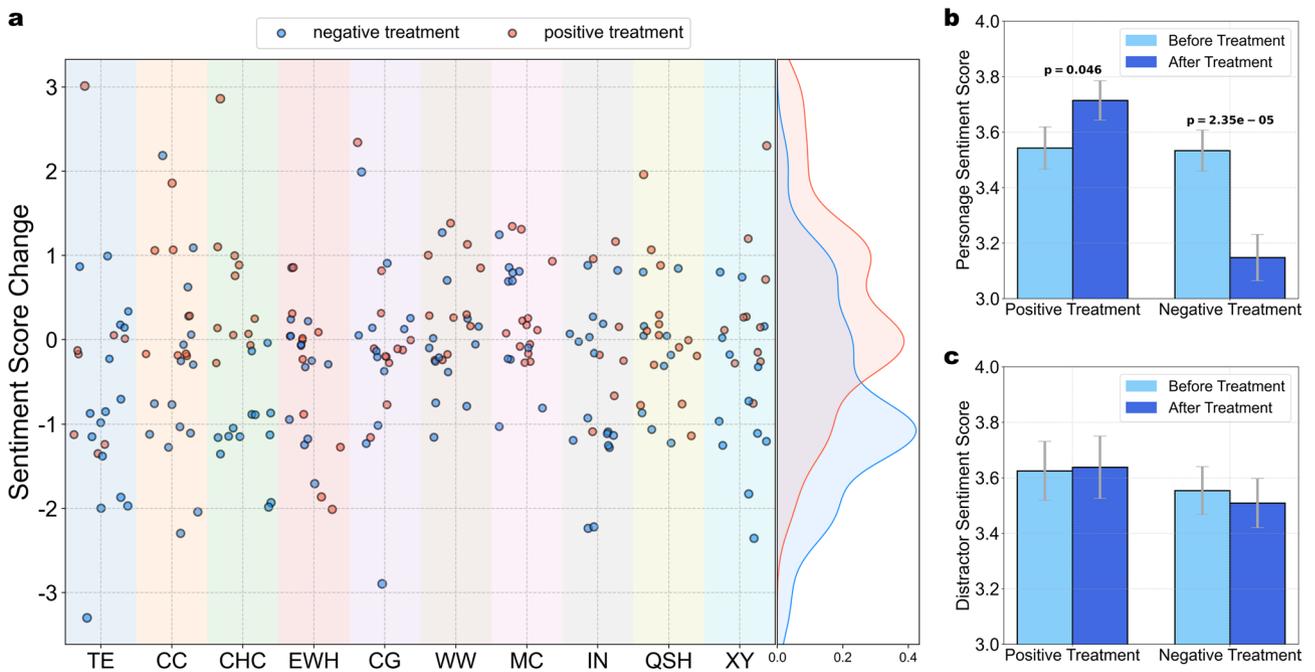
After exposure to positive content, the average values of valence while watching negative videos significantly increased ( $Mean_{diff} = 0.353$ ,  $w = 858.0$ ,  $p = 0.043$ ,  $r = 0.275$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ). Concurrently, the arousal while watching negative videos



**Fig. 1.** The overall procedure of the three-stage user studies.  $S_1$ : Participants browsed videos featuring positive or negative polarity on 10 personages. Their behavior signals, brain signals, and explicit responses are collected during the video browsing process or the post-experiment questionnaire.  $S_2$ : Participants browsed short videos about these personages posted by the Platform in a field study lasting 6 days. For each personage, the Platform only recommended videos with either positive or negative polarity, which acts as a manifestation of CP.  $S_3$ : The last stage resembles the procedures in stage  $S_1$ . However, it diverges in the selected videos and the timing of the questionnaire.



**Fig. 2.** LAB Study: Experimental Setup and Interface. (a) The LAB study experiment primarily involves participants watching videos on a monitor, some of which contain positive or negative emotions toward a character, while their EEG signals are recorded using an EEG cap. Participants adjust to a comfortable sitting position in advance and strive to minimize vigorous physical movements to reduce interference with the EEG signals. The EEG cap is a non-invasive device that does not interfere with the participants’ normal video-viewing experience. (b) A screenshot of the experimental platform’s webpage shows the video playback window in the center with a “like” button below. After the video has played more than halfway, participants can click the “nextone” button at any time to proceed to the next video. Each participant has an independent account.



**Fig. 3.** Overview of sentiment scores and their changes before and after different treatments. (a) Scatter plot of score changes for ten selected personages (i.e., TE, CC, CHC, EWH, CG, WW, MC, IN, QSH, XY, details are introduced in SI Appendix, Section 5) before and after exposure to positive (red) and negative (blue) content. The frequency distribution illustrates the distribution of score changes under positive and negative treatments. (b) Average sentiment scores across participants and personages before and after exposure to polarized content (positive and negative). Positive treatment significantly increases personage scores ( $p=0.046$ ), while negative treatment significantly decreases them ( $p=2.35e-5$ ). (c) Average sentiment scores across participants and distractors before and after exposure to positive and negative content. Distractors refer to entities that do not contain any relevant videos in the field study. Neither treatment significantly impacts distractor scores.

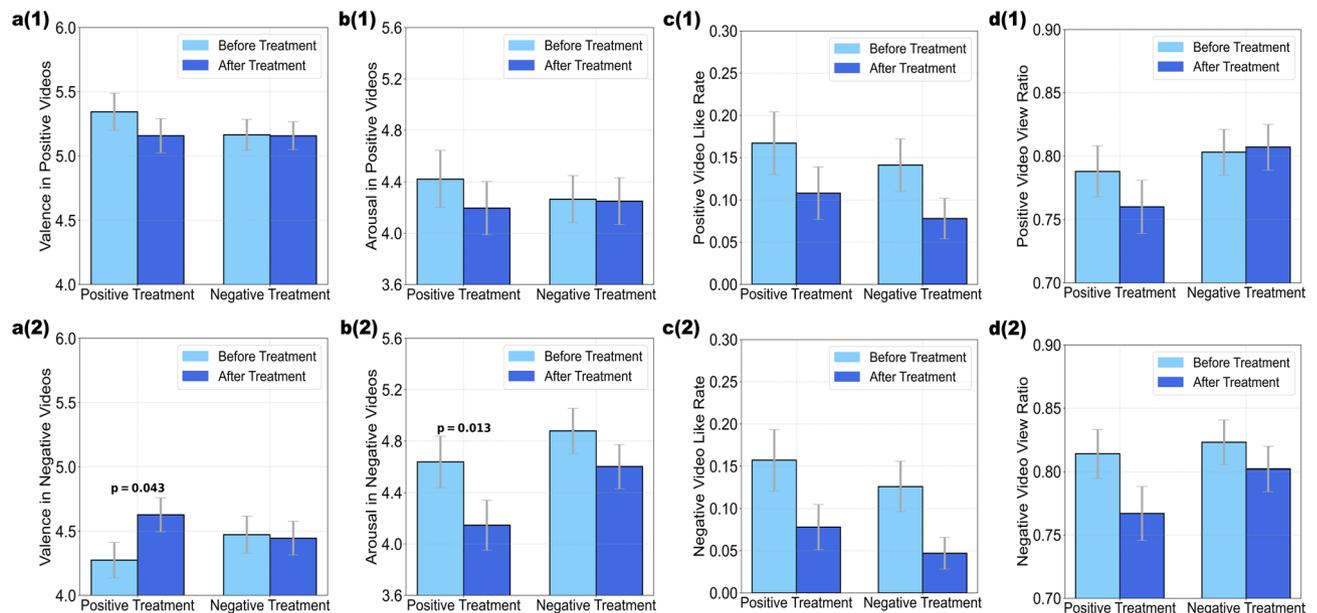
significantly decreased ( $Mean_{diff} = -0.490$ ,  $w = 1065.5$ ,  $p = 0.013$ ,  $r = 0.303$ , see Fig. 4). In contrast, after exposure to negative content, the like rate, valence, arousal, and view ratio slightly decreased ( $Mean_{diff} = -0.078$ ,  $-0.027$ ,  $-0.277$ , and  $-0.022$ ,  $p > 0.05$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ). Exposure to the class of polarized content (i.e., positive or negative) is significantly correlated with participants' brain responses to videos. Specifically, in the delta band, significant correlations were observed at electrodes FT7 (Pearson's  $r = 0.1281$ ,  $p = 0.0073$ , band power difference = 0.0936), FC5 (Pearson's  $r = 0.1067$ ,  $p = 0.0258$ , band power difference = 0.0608), and C5 (Pearson's  $r = 0.1071$ ,  $p = 0.0252$ , band power difference = 0.0805). For the gamma band, a significant correlation was found at electrode FT7 (Pearson's  $r = -0.0971$ ,  $p = 0.0425$ , band power difference =  $-0.2049$ , Fig. 5 b). The correlation between content polarization and brain responses, grouped by varying valences such as rate and view ratio, is observed in the delta and theta frequency bands across the prefrontal, frontal, and parietal lobes (see Fig. 5 c-f).

### Sentiment annotations before and after polarized video treatment

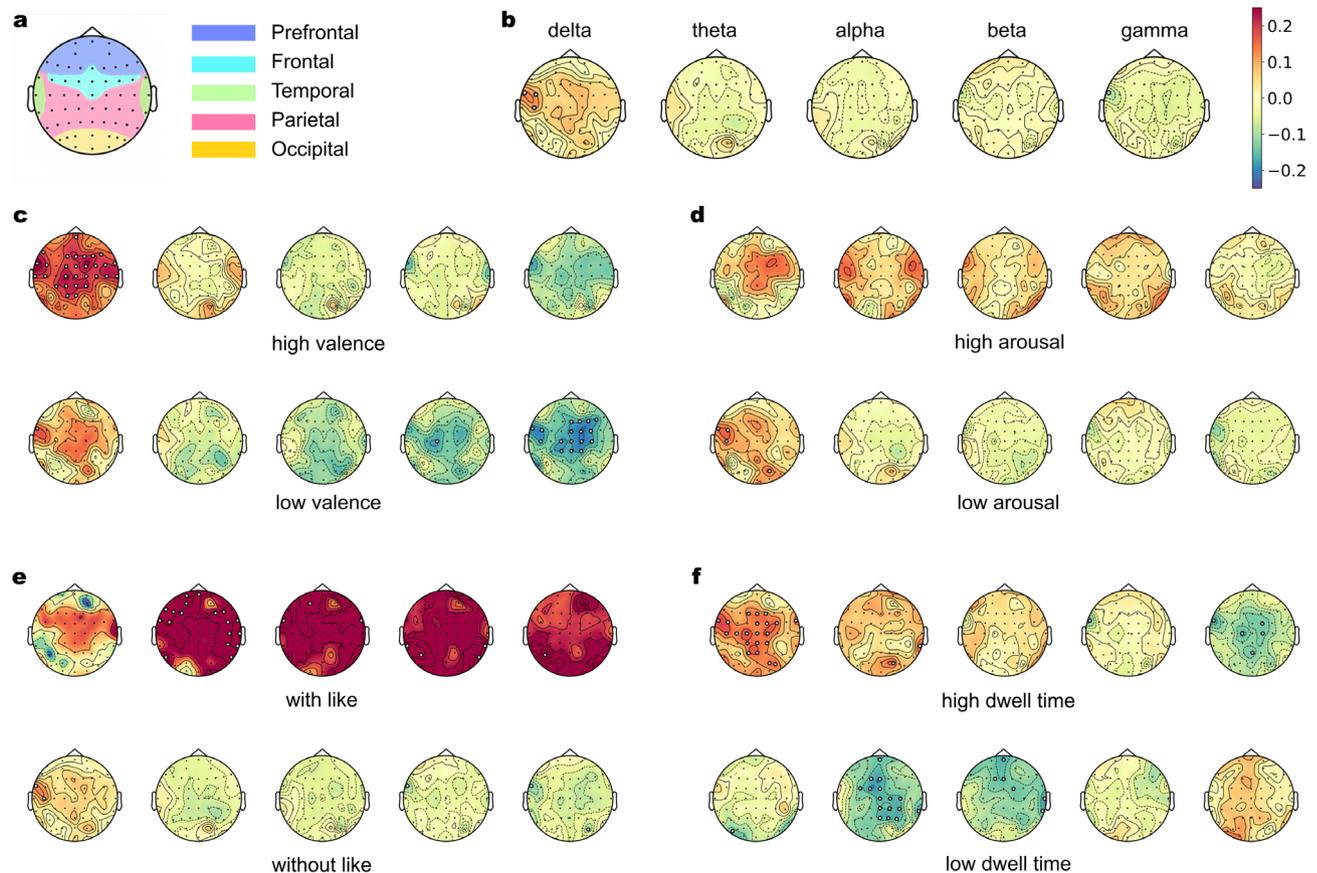
To examine the influence of content polarization (CP) on user sentiment, we collected sentiment scores for each personage from participants before (i.e., Fig. 1 S1) and after (i.e., Fig. 1 S3) the field study using a five-point Likert scale. After engaging with positive treatment videos, we can observe a significant upward shift in emotions. This indicates that participants are influenced by the opinions in the short videos and therefore exhibit a more positive sentiment toward these figures (see Fig. 3a). In contrast, after negative treatment, there is a substantial downward shift in emotions. On the other hand, the users' sentiment scores for the distractors were also collected to serve as a control in our study. Distractors refer to entities that do not contain any relevant videos in the field study. Fig. 3b, c show that before exposure to the experimental conditions, there was no significant difference in average sentiment scores between the groups (3.543 and 3.533). Sentiment scores increased significantly in the positive treatment, from 3.543 (Pre Positive) to 3.714 (Post Positive). After the negative treatment, scores decreased notably from 3.533 (Pre Negative) to 3.148 (Post Negative). The sentiment scores for distractors showed no significant change, as shown in Fig. 3c. These results indicate that the polarized content could significantly affect users' sentiment judgment towards personages.

### Emotional states before and after polarized video treatment

We collected participants' emotional states (i.e., valence and arousal) while watching each video during both the Pre-study and Post-study phases. The emotional states were measured using a Likert scale ranging from 1 to 9, reflecting participants' emotions while browsing the videos.



**Fig. 4.** Scatter plots of the averaged valence and arousal scores, like rates, and view ratios for viewing positive and negative content videos about the personages before and after exposure to content with positive or negative opinions. **(a(1), a(2)) Averaged Valence Scores** for viewing positive (a(1)) and negative (a(2)) content videos, measured before and after exposure to content with positive or negative opinions. **(b(1), b(2)) Averaged Arousal Scores** for viewing positive (b(1)) and negative (b(2)) content videos, measured before and after exposure to content with positive or negative opinions. **(c(1), c(2)) Averaged Like Rates** for videos with positive opinions (c(1)) and negative opinions (c(2)), recorded before and after exposure to content with positive or negative opinions. **(d(1), d(2)) Averaged View Ratios** for videos with positive opinion (d(1)) and negative opinion (d(2)), calculated as the sum of viewing times divided by the total video duration and count, measured before and after exposure to content with positive or negative opinions.



**Fig. 5.** (a) Brain region location diagram: Each color corresponds to a brain region, and the dots represent electrode positions. (b) Pearson correlations between polarized videos and differential entropy (DE) in EEG frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz) during the laboratory study. A white circle represents a significant correlation ( $p < 0.05$ ) between DE features and content polarization. (c, d, e, f) The Pearson correlations of the opinion polarization with DE in frequency bands of delta, theta, alpha, beta, and gamma grouped by various factors in the lab study: c, valence (positive or negative mood); (d) arousal (high or low arousal mood); (e) like data (like or view without like); and (f) view ratio (proportion of video viewed, long or short).

Fig. 4a(1), a(2) show the influence on valence while watching either positive or negative opinion videos. Generally, videos with positive opinion (refer to Fig. 4a(1)) have higher valence scores compared to negative opinion videos (see Fig. 4b(1)) ( $Mean_{diff} = 0.745$ ,  $w=12905.0$ ,  $p=2.67e^{-18}$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ). This indicates that content presenting positive opinions generally evokes higher valence. When comparing the difference in the valence score between the pre-study and the post-study, we observe that the positive treatment had a significant impact on valence while watching videos featuring the same personage. Specifically, valence scores showed a notable decrease from 5.333 to 5.152 when viewing positive opinion videos and a significant increase ( $p=0.043$ ,  $r=0.275$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ) from 4.238 to 4.619 for negative opinion videos.

In terms of arousal, exposure to both positive and negative treatment results in a decrease (see Fig. 4b(1), b(2)). This decrease is more pronounced in a positive treatment, with arousal levels falling from 4.422 to 4.196 for positive videos and from 4.637 to 4.147 for negative ones ( $p=0.013$ ,  $r=0.303$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ). Conversely, in a negative treatment, the arousal reduction is less marked, decreasing from 4.266 to 4.250 for positive videos and from 4.879 to 4.602 for negative ones ( $p>0.05$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ). A partial inference from this observation is that positive treatment may result in greater information fatigue conveyed by the short videos. Additionally, arousal levels for negative opinion videos (Fig. 4b(1)) are generally higher than that of positive ones (Fig. 4b(2)). This indicates that videos with negative opinions more effectively capture immediate user attention.

### Behavior response before and after exposure to polarized content

We analyzed the impact on participants' behavioral data, including likes and view ratios. The behavioral data were collected before and after the treatment, as shown in Fig. 1 S1, S3. The like rate and view ratio are calculated as follows:

$$\text{Like Rate} = \frac{\# \text{ Liked videos}}{\# \text{ Videos}}, \quad \text{View Ratio} = \frac{\text{Total Viewing Time}}{\# \text{ Videos} \times \text{Duration per Video}} \quad (1)$$

Fig. 4c(1), c(2) illustrates a significant decline in the like rate when participants watched either positive videos or negative videos after they were exposed to either positive or negative treatments. This indicates a diminished user interest in monothematic videos. A comparison of the impacts of polarized content revealed no significant difference in the like rates between positive and negative influences ( $Mean_{diff} = 0.028$ ,  $p > 0.05$ , two-sided paired Wilcoxon signed-rank test due to non-normal distribution, Shapiro-Wilk  $p < 0.001$ ). This might suggest that like rates are not satisfactory metrics in probing the impact of polarized content on users.

As shown in Fig. 4d(1), d(2), the view ratio, averaged for positive and negative polarity videos, ranges from 0.5 to 1.0, as the users are allowed to skip a video after 0.5 minutes and the total duration of each video is 1 minute. From Fig. 4d(1), d(2), we observe that following exposure to a positive opinion treatment, the view ratio for positive opinion videos showed a marginal decrease, which was not statistically significant according to a two-sided paired Wilcoxon signed-rank test ( $Mean_{diff} = 0.004$ ). In contrast, for negative opinion videos, the decrease was significant as confirmed by the same test ( $Mean_{diff} = 0.022$ ,  $p = 0.039$ ). In summary, relative to sentiment annotations, the view ratio did not exhibit a notable difference in the influence of positive versus negative content on users.

**Answer to RQ1:** Sentiment score analysis reveals that content polarization significantly influences user sentiments, making them more aligned with the prevailing opinions of the content. On the other hand, the user behavior data (like and view ratio) has a relatively weaker association with polarized content. This implies that users' sentiment judgments and emotion annotation potentially offer a more accurate reflection of the impact of CP than conventional behavioral metrics.

### Neural correlates of content polarization in EEG

To investigate the relationship between EEG signals and opinion polarization, we calculated the Pearson correlation coefficients between the Differential Entropy (DE) (see Method Feature Extraction) feature of EEG and the CP. This relationship is assessed by correlation between CP treatments (positive or negative) in the field study and the EEG signals collected during the post-study. For each frequency band (delta, theta, alpha, beta, and gamma) and each electrode, we first calculated the average DE feature across the 30-second time window for each user-personage combination. We then computed the Pearson correlation coefficient between these electrode-specific, frequency-band-specific DE values and their corresponding content polarity labels<sup>33,34</sup>. It should be noted that the FPz electrode did not show significant correlations with content polarization in this analysis ( $p > 0.05$  across all frequency bands), despite its utility in our prediction model. To account for multiple comparisons across electrodes and frequency bands, we employed cluster-based permutation testing<sup>35</sup>, which preserves the spatial relationships among electrodes while controlling for Type I errors. Unlike traditional Bonferroni or FDR corrections that treat each comparison as independent, this approach acknowledges the spatial dependencies in EEG data where adjacent electrodes often capture related neural activity. The cluster-based analysis revealed significant correlations in the Delta frequency band in the Left Temporal region (electrodes FT7, T7, TP7;  $t=2.69$ ,  $p=0.017$ ), with trending significance in the Delta band Frontal region (electrodes FC5 et al;  $t=2.24$ ,  $p=0.083$ ) and Gamma band Left Temporal region (electrodes FT7, T7, TP7;  $t=2.03$ ,  $p=0.082$ ). These findings largely correspond with the electrode-specific results presented in Fig. 5. While we report the electrode-wise topographic maps to maintain spatial precision of the electrode-level effects, the statistical robustness of our findings is supported by the cluster-based permutation results (see SI Appendix Fig.s9 for detailed cluster-based analysis). This analysis was performed separately for each frequency band and electrode, resulting in the correlation patterns shown in Fig. 5.

Fig. 5 (b-f) presents the correlations between Content Polarization and brain responses. In the delta frequency band, a positive correlation exists between positive CP treatments and EEG signals, suggesting that positive treatments may enhance low-level EEG activity in the Delta band (see Fig. 5(b)). Conversely, negative CP treatments are associated with a relatively negative correlation in the Delta band with EEG signals. We observed that positive CP treatments generally exhibit a strong negative correlation with EEG signals in the gamma band.

### Neural correlates of opinion polarization in EEG with emotional states

Given that brain signals are susceptible to interference from the subject's current emotional state, we categorized brain signals based on emotional states for correlation analysis. Fig. 5 c illustrates that, in both high and low valence, there is a significant correlation between CP types and brainwave signals. In high valence scenarios, brain activity in terms of DE in the delta frequency band, particularly in the prefrontal, frontal, and parietal regions, is significantly positively correlated with experiencing positive short video treatment. In low valence scenarios, brain activity (higher band power) in the gamma frequency band, linked to cognition and psychological experiences, particularly in the frontal and parietal regions, is significantly negatively correlated with experiencing positive short video treatment. EEG signals were categorized based on arousal, as illustrated in Fig. 5 d. In addition, a comparison of results under high and low arousal conditions revealed a stronger positive correlation between CP and brainwave signals in the theta and gamma frequency bands under high arousal states.

### Neural correlates of content polarization in EEG with behaviors

To explore the impact of behavioral data on the correlation between CP and EEG, brainwave signals were categorized based on 'likes' received and view ratio. Videos were classified into two categories: those with a view

Input Features				ACC	F1	AUC
EEG	Like	Emotion	Sentiment	Mean $\pm$ SEM	Mean $\pm$ SEM	Mean $\pm$ SEM
-	-	-	✓	0.592* $\pm$ 0.029	0.588* $\pm$ 0.028	0.662 $\pm$ 0.038
-	-	✓	-	0.517* $\pm$ 0.041	0.517* $\pm$ 0.040	0.504* $\pm$ 0.033
-	-	✓	✓	0.497* $\pm$ 0.030	0.497* $\pm$ 0.029	0.530* $\pm$ 0.026
-	✓	-	-	0.557* $\pm$ 0.015	0.485* $\pm$ 0.021	0.548* $\pm$ 0.023
-	✓	-	✓	0.607 $\pm$ 0.023	0.598 $\pm$ 0.023	0.630 $\pm$ 0.034
-	✓	✓	-	0.552* $\pm$ 0.028	0.550* $\pm$ 0.027	0.514* $\pm$ 0.032
-	✓	✓	✓	0.537* $\pm$ 0.014	0.536* $\pm$ 0.014	0.581* $\pm$ 0.025
✓	-	-	-	0.641 $\pm$ 0.024	0.639 $\pm$ 0.024	0.672 $\pm$ 0.029
✓	-	-	✓	0.622 $\pm$ 0.019	0.620 $\pm$ 0.019	0.677 $\pm$ 0.016
✓	-	✓	-	0.597* $\pm$ 0.017	0.594* $\pm$ 0.017	0.639 $\pm$ 0.025
✓	-	✓	✓	0.631 $\pm$ 0.025	0.629 $\pm$ 0.024	0.672 $\pm$ 0.019
✓	✓	-	-	0.627 $\pm$ 0.022	0.623 $\pm$ 0.023	0.670 $\pm$ 0.030
✓	✓	-	✓	<b>0.647</b> $\pm$ 0.019	<b>0.644</b> $\pm$ 0.020	<b>0.687</b> $\pm$ 0.021
✓	✓	✓	-	0.577* $\pm$ 0.026	0.572* $\pm$ 0.024	0.634 $\pm$ 0.024
✓	✓	✓	✓	0.646 $\pm$ 0.016	<b>0.644</b> $\pm$ 0.016	0.685 $\pm$ 0.027

**Table 1.** The performance of predicting content polarity using different input features and their combinations, both before and after adding EEG data. Adding EEG data significantly improved ACC, F1, and AUC for all feature combinations. When using the combination of EEG, Behavior, Emotion, and Sentiment features, the model achieved an ACC of 0.646, an F1 score of 0.644, and an AUC of 0.685.

ratio above 0.9 as a long view ratio (high time), and those below 0.9 as a short view ratio (low time). This method yielded two balanced data sets, with an approximate ratio of low view count to high view count sets being 10:9.

Fig. 5 clearly illustrates a significant increase in the positive correlation between CP and EEG signals when grouped by 'like', specifically observed across the different frequency bands. For videos receiving likes in post-study, stronger EEG signals across multiple frequency bands are observed after positive treatment. Significantly, correlations between CP and brain responses were found in the theta and alpha bands within the prefrontal and frontal lobe regions, an area linked to the generation and processing of emotions<sup>36</sup>.

As shown in Fig. 5 f, the data were grouped into two categories based on the length of the view ratio. In the long viewing ratio group, a majority of electrodes in the delta band, mainly in frontal and temporal regions, showed a notable positive correlation, while the gamma band showed significant negative correlations. For the short viewing ratio, significant negative correlations were observed in the theta band, especially in the prefrontal, frontal, and parietal regions. Significant negative correlations were also observed in the prefrontal region in the alpha band.

**Answer to RQ2:** Our study found that EEG signals and exposure to CP are intricately linked, with significant correlations in the Delta and Gamma frequency bands. Positive CP is associated with increased Delta band activity. In contrast, negative polarization correlates with increased activity in the Gamma band, potentially indicating more active memory recall and emotional responses<sup>37,38</sup>. We also noted that valence, rather than arousal, is a stronger interfering factor in CP types. Furthermore, behavioral data (i.e., 'likes' and view ratio) also act as interfering factors, and their patterns differ from those of emotional data. Videos receiving more likes or with a longer view ratio show strong correlations between CP and brain responses, indicating that higher user engagement enables potentially more effective detection of exposure to CP with EEG inputs.

### CP prediction model

Motivated by the observed correlation between CP and user signals, we further investigate to what extent CP exposure can be detected with those signals and address RQ3. We categorize the content browsed by users in the field study into two categories: positive and negative. The classification results are detailed in Table 1, evaluated using metrics such as the Area Under the Curve (AUC), F1 Score (F1), and Accuracy (ACC)<sup>39</sup>.

First, when relying solely on individual behavioral metrics (i.e., like or view ratio), the prediction performance was notably weak, with AUCs only reaching 0.548 (like) and 0.520 (view ratio). Combining likes and view ratio features unexpectedly led to a decline in performance. This further verified our previous analysis and showed that relying on behavior measurements is not effective in CP detection. Second, **using EEG signals as a standalone feature yielded superior predictive results compared to traditional behavioral features.** The performance metrics indicated considerable improvements over behavioral-based predictions, demonstrating EEG's effectiveness in predicting users' exposure to polarized short videos. Furthermore, when sentiment annotations were independently used as features, they showed enhanced performance compared to behavior-based models. Although the AUC performance of sentiment annotations was close to that of EEG, EEG achieved significantly higher F1 and ACC metrics. EEG not only outperformed behavioral features but also surpassed sentiment or emotion features in terms of AUC, F1, and ACC metrics. Further, combining EEG and behavior features surpassing the results obtained using only behavior features in ACC (from 0.557 to 0.627,  $p=0.040$ ), F1 (from 0.485 to 0.623,  $p=0.032$ ), and AUC (from 0.548 to 0.670,  $p=0.018$ , pair-wise t-test). Combining EEG and sentiment features also showed better performance than using sentiment alone. Moreover, the combination

of behavior and sentiment produced better performance in AUC (0.630) than using behavior (0.548,  $p=0.022$ , pair-wise t-test) alone.

Last, **a combination of behavior, sentiment, and EEG features led to the best performance, with an ACC of 0.647, an F1 score of 0.644, and an AUC of 0.687.** The high performance reflects a significant likelihood of accurately predicting the exposure to CP. Combined results demonstrate that EEG features substantially contribute to predicting exposure to CP.

**Answer to RQ3:** The experimental findings in the polarity classification task indicate that EEG signals, along with explicit responses (i.e., sentiment annotations, valence annotations, and behavior responses), can be effectively utilized to detect participants' exposure to polarized content. This method demonstrates superior performance compared to relying on traditional behavioral information.

## Discussions

The participants' explicit feedback regarding their sentiments toward specific personages, as reflected in the pre- and post-treatment survey results, underwent significant changes. This indicates that exposure to positive and negative content indeed significantly impacts users' emotional tendencies toward the personage. Based on the differences in significance (positive treatment  $p=0.038$ , negative treatment  $p=5.62e-06$ ), we speculate that negative treatment (short videos with negative attitudes) has a greater impact on user sentiment. Under the same conditions, we infer that content involving criticism and blame has a stronger impact on users' emotions than content involving praise.

In addition to assessing direct sentiment towards specific personages, we evaluated users' emotional (valence, arousal) and behavioral (like rate, view ratio) responses to videos themed around particular personages. We found that regardless of the type of treatment, participants exhibited significantly higher average valence when watching positive videos compared to negative videos. Therefore, we speculate that, under the same conditions, content involving praise is more likely to elicit pleasure than content involving criticism and blame, regardless of the participants' attitudes towards the content itself. A somewhat counterintuitive finding is that after experiencing positive treatment, that is, watching a large number of praising videos, participants showed a significant increase in valence when exposed to critical content. This suggests that excessive praise can easily lead to information fatigue, resulting in a rebound effect. We reveal the differing impacts of positive and negative polarized content on users, highlighting that positive content has a greater influence on users' emotional and behavioral responses. Specifically, positive treatment significantly increased valence and markedly reduced arousal, like rate, and view ratio when users watched negative videos of the same personage. In contrast, negative treatment only significantly reduced the like rate. These findings suggest that the impact of positive treatment on emotions and behaviors is more pronounced than negative treatment.

More intriguingly, content polarization has a significant impact on users' brain activity, as measured by EEG devices. Specifically, after undergoing a certain type of treatment, changes in brain activity occur when users are re-exposed to the same topic content, particularly in specific brain regions (the frontal and prefrontal lobes) and EEG frequency bands. Our results showed that the activity in the delta band was positively correlated with positive CP treatments, while the gamma band showed a negative correlation with such treatments. Delta waves have been associated with various states and processes relevant to content processing, including unconscious information processing<sup>40</sup>, emotional regulation<sup>41</sup>, reward processing<sup>42</sup>, and certain forms of attention to internal stimuli<sup>43</sup>. One possible interpretation of our findings is that the increased delta activity following positive content exposure might reflect changes in attentional processes<sup>44</sup>, emotional regulation<sup>45</sup>, or memory consolidation<sup>46</sup>. At the neurophysiological level, this may involve specialized information transfer mechanisms between the limbic system and frontal cortex. Research suggests that structures such as the hippocampus and amygdala communicate with the prefrontal cortex through rhythmic delta oscillations, creating temporal windows for emotional information integration into higher cognitive processing<sup>41</sup>. When users encounter positively polarized content that aligns with their beliefs, this may enhance limbic-frontal connectivity via delta-band synchronization, potentially reinforcing both emotional experiences and related belief systems. However, alternative explanations cannot be ruled out, and the precise relationship between delta activity and CP requires further investigation. The gamma band has been recognized as crucial for learning and memory processes<sup>47</sup> and has also been correlated with meditation<sup>48</sup>. The observed correlation between gamma activity and negative content exposure might potentially be related to enhanced cognitive processing or emotional salience of such content. This relationship may reflect the dynamic interaction between the amygdala and prefrontal cortex during processing of challenging or emotionally negative information. When confronted with such content, the amygdala rapidly activates and signals emotional salience through gamma synchronization, while prefrontal regions simultaneously engage inhibitory control mechanisms to regulate these emotional responses<sup>49,50</sup>. The negative correlation we observed could indicate enhanced prefrontal regulatory control over amygdala-driven emotional responses when processing challenging information, implemented through alterations in local gamma oscillation patterns. However, it is important to note that gamma oscillations serve multiple functions in information processing, and their relationship with content valence is complex and still being investigated<sup>51</sup>.

A comparison between positive and negative mood states revealed a more substantial positive correlation in the delta band under positive mood. These findings suggest that emotional valence may modulate the relationship between content polarization and neural oscillations. While delta activity has traditionally been associated with certain unconscious processes, it is also involved in reward processing and emotional regulation<sup>41</sup>, which could potentially explain its relationship with positive mood states in our study. It was observed that correlations derived from arousal-based grouping were weaker than those related to valence. This finding corroborates the analysis regarding valence and arousal sentiment scores, suggesting that valence is a stronger interfering factor in CP types than arousal. Therefore, in the detection of content polarization or tendencies, particularly when using

EEG paradigms, we speculate that controlling valence, an important interfering factor, can enhance the accuracy of detecting and understanding users' real-time emotional responses.

The "like" and view ratio are fundamental behavioral data collected by internet platforms. After filtering using "likes," we observed the highest range of positive correlations. We speculate that liking a video indicates high user engagement and agreement with the video's perspective. Therefore, high engagement and agreement with the content significantly enhance the brain's ability to reflect the impact of content polarization (CP). The view ratio is also a valuable metric. For videos watched more completely, higher band power was observed in the delta frequency band following positive treatment. This implies that more complete engagement with a video may result in increased unconscious brain activity associated with positive content. For videos that users skipped quickly, weaker activity was observed in the theta frequency band in the same brain regions following positive treatment. This implies that less complete engagement with a video may result in lower attention and cognitive control associated with positive content. In summary, both "like" and view ratio can interfere with the analysis of CP as well as the collected EEG responses. Therefore, in future CP detection on internet platforms or algorithms, such behavioral log data can be used for grouping to more accurately analyze the extent and impact of CP.

In our predictive experiments, we explored the feasibility of using electroencephalography (EEG) signals to predict the polarity (positive/negative) of users' Cognitive Polarity (CP), which reveals potential practical implications of previously discovered correlations and characteristics of specific brain regions.

EEG signals provide a more precise and insightful measure of users' implicit responses to polarized content, uncovering impacts that traditional behavioral data might overlook. On the one hand, the results of our research, especially our prediction model, may be a step towards developing more robust methods of early detection of online opinion polarization about socially significant topics, contributing to better social media governance. On the other hand, our study develops methods to evaluate how interactions with polarized social media material affect individuals on the physiological level, which further studies in the areas of psychology, sociology, computer science, and media and communication may employ. With technological advancements in portable brain data collection methods, brain-signal-based devices can be integrated into Information Retrieval (IR), virtual reality, services for the disabled, and Natural Language Processing (NLP) applications such as intelligent assistants, enhancing accuracy and offering real-time feedback in these areas.

One potential limitation of our study is that our participants were Asian adults aged 18 to 25. This represents an important demographic of social media users, but future studies could benefit from including participants across a wider age spectrum. Additionally, although most EEG-related studies do not focus on racial balance, we plan to collaborate with research groups in other countries in the future to expand the cultural diversity of the participant pool to enhance the generalizability of findings. Although this study's sample size was comparable to those typically used in EEG research, future studies could use our proposed paradigm to conduct experiments in various environments. Furthermore, despite our efforts to increase realism by designing custom web pages and incorporating background videos, there remains a discrepancy between the video content and users' daily experiences. Future research should consider how to more effectively integrate everyday life experiences into video content to boost the study's practicality and relevance. Additionally, while our current prediction approach using XGBoost with EEG features combined with sentiment and behavioral data showed promising results, several more sophisticated modeling approaches could further improve prediction accuracy and interpretability. Specifically, we envision four promising directions: (1) Temporal Transformers for EEG time-series analysis, (2) multimodal fusion networks with attention mechanisms, (3) Graph Neural Networks for brain connectivity modeling, and (4) neuro-symbolic systems incorporating domain knowledge. These advanced approaches would likely enhance our ability to detect and understand neural responses to polarized content by modeling complex temporal dynamics and non-linear relationships in the data.

## Methods

In this section, we first present how we selected target personages, constructed a corresponding video set, and designed a video platform (SI Appendix, Section 3, Website Design Specification) for the experiment. Then, we introduce the whole experiment's procedure, including the participants' recruitment details. We also introduce experimental methods for predicting content polarity.

### User study design

#### *Target personages selection*

To study how short video browsing affects people's opinions, we selected ten historical and fictional personages based on the following criteria: (1) The personages should be such that people may hold differing opinions about them (2) The personages should be non-modern figures to avoid ethical issues. Specifically, based on the background of the participants (mainly Asians), we selected personages including Thomas Alva Edison, Cao Cao, Christopher Columbus, Emperor Wu of Han, Catherine the Great, Walter White, Michael Corleone, Isaac Newton, Qin Shi Huang, and Xiang Yu. The introductions to these personages are shown in the SI Appendix, Section 4, and Table. S6. We investigated the familiarity of the participants with these personages through questionnaires (SI Appendix, Section 2) in the pre-study and post-study. According to the pre-study questionnaires, participants had a moderate level of familiarity with these personages, averaging a score of 2.8 on a Likert scale ranging from 1 to 5. After our experiments, this average familiarity score increased to 3.2. The detailed distribution of these scores is shown in the SI Appendix, Fig. S1. This indicates that the participants feel they acquired relevant information about the personages during the experiment.

#### *Video set construction*

We constructed a video set that included the polarized content related to the ten personages to conduct experiments simulating short video browsing by the participants. The video set construction involved video

collection and annotation processes. We collected authentic, real-world videos from two popular video platforms, i.e., Tiktok (<https://www.douyin.com/>) and bilibili (<https://www.bilibili.com/>) by submitting the target personages as queries. The criteria for selecting videos are (1) the main topic of the video should be the target personage, and (2) the video should exhibit clear polarity (i.e., positive or negative attitude to the personage). As videos of unequal lengths could have different effects on viewers, we truncated each video to a duration of one minute while preserving the key content of the video. We collected 16 videos for each personage, amounting to 160 videos in total.

We recruited three annotators: two Ph.D. candidates in Computer Science and one undergraduate student in Journalism and Communication. They all have extensive experience in sentiment annotation. Before starting the annotation tasks, all annotators completed a thorough training program (SI Appendix, Section 1) to ensure a standardized understanding specific to our task. Each video was annotated, ranging from 1 to 5, following a Likert scale, where 1 represents the most negative opinion, while 5 signifies the most positive opinion. The polarity of the videos was determined based on the average scores from three experts, where scores above and below 3 denote positive and negative polarity, respectively.

To better simulate real-world short video browsing scenarios and prevent the participants from identifying the goal of our lab and field studies, we also included a background video set for the field study. This set contains 120 one-minute clips unrelated to the target personages, also obtained directly from TikTok and Bilibili, each also edited to one minute in duration without altering the original content. Additionally, 40 videos from the SEED-IV dataset (a well-known dataset for sentiment analysis)<sup>52</sup>, were included and truncated to a one-minute duration. The final video set comprised four types of videos: **videos about target personages with positive opinions (pos-videos)**, **videos about target personages with negative opinions (neg-videos)**, **background videos**, and **SEED videos**.

### Platform

To facilitate data collection, we developed a custom experimental webpage for video playback. The webpage mimics standard short video platforms where participants can watch short videos in a cascade sequence. At the same time, the webpage can collect user interactions, including “likes” and viewing duration, mimicking standard short video platforms. To accommodate diverse short video browsing scenarios, the webpage has been optimized for display on both mobile phone and laptop screens. As shown in Fig. 2, the webpage’s playback interface primarily features a video frame, a “Next One” button to skip to the next video, a “Last One” button to go back to the previous video, and a heart button for “likes”. We provided an individual account to each participant, and they had to log in before they started browsing the video. To validate the ecological validity of our custom platform, we compared user engagement metrics with publicly available statistics from commercial platforms. The results showed striking similarities: users viewed videos on our platform for 43.0% of their total duration (compared to 44% on TikTok) and had a like rate of 12.2% (compared to TikTok’s 11%)<sup>53</sup>. These similarities suggest participants’ interactions closely approximated their natural browsing habits on commercial platforms, supporting our experimental design’s ecological validity.

### Procedure

The procedure of our user study is summarized in Fig. 1. It consists of three stages: pre-study, field study, and post-study. In the **pre-study**, participants were trained on the video browsing platform, watched a randomized set of videos while EEG signals were recorded, and completed a questionnaire to provide sentiment and familiarity ratings for personages. Distractor questions were included to minimize bias. The **field study** exposed participants to daily short videos over six days, with systematic content bias (positive or negative) applied to specific personages while ensuring a balanced and randomized video environment. Daily questions were used to confirm participant engagement. Finally, in the **post-study**, participants repeated the EEG experiment and questionnaire, with the sequence adjusted to isolate the effects of the field study. Please refer to SI Appendix, Section 8 for further details.

### Participants and ethics

Our experiment recruited 24 participants, aged 18 to 25, identified as active users of at least one short video platform based on a pre-experiment survey (SI Appendix, Fig.S8). Our participants exhibit diverse usage patterns across multiple platforms (SI Appendix, Fig. 11), varied content preferences (SI Appendix, Fig. 12), and a range of cognitive processing styles (SI Appendix, Fig. 13). Prior to the main experiment, we conducted a pilot study ( $n = 3$ ) to estimate the required sample size. A Wilcoxon signed-rank test (used due to non-normal distribution) revealed a significant difference between positive and negative environmental effects with a large effect size (Cohen’s  $d = 2.17$ ). Based on this effect size, we performed a power analysis with parameters: significance level  $\alpha = 0.05$  and statistical power = 0.90 (two-sided test). The minimum required sample size was calculated using:

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \times 2}{d^2} \quad (2)$$

where  $Z_{\alpha/2}$  represents the critical value of the normal distribution at  $\alpha/2$ ,  $Z_{\beta}$  is the critical value at  $\beta$  (1 - power), and  $d$  is the standardized effect size. This analysis indicated a minimum requirement of 12 participants to reliably detect the expected effects. We recruited 24 participants, exceeding the minimum requirement of 12 and ensuring sufficient statistical power for detecting the anticipated effects. The final sample comprised 13 males and 11 females from various majors, including Computer Science, Environmental Science, Automation, and Life Sciences. All participants are right-handed and reported no history of neurological disorders, ensuring uniformity in terms of brain function and manual dexterity for the tasks involved in the study. Note that, after

the whole experiment procedure, we explicitly informed the participants about our experiment objectives and let them decide whether we could use their data for future analysis to avoid potential ethical risks.

The collected dataset consists of 24 participants, among which one participant's data was discarded for technical issues, yielding data for 230 user-personage pairs. Each user-personage pair includes EEG signals, behavior responses (i.e., like and duration) collected while viewing the video about the specific personage, and their attitudes to the personage collected through questionnaires.

This study was reviewed and approved by the Department of Science and Technology Ethics Committee, Tsinghua University (THU01-20230221). All experiments were performed in accordance with relevant guidelines and regulations. We ensured that the experiments were harmless to all participants. Participants were required to sign an informed consent form (SI Appendix, Section 5) before the user study. The consent form detailed the nature and purpose of the study and emphasized the confidentiality and anonymity of responses. It also assured participants of their right to withdraw from the study at any point without any penalty.

## Content polarization prediction

### *Task formalization*

To investigate exposure to CP in the context of specific topics or personages, it is essential to have a reliable method for determining the direction of opinion polarization, i.e., positive or negative opinions towards the personages<sup>54</sup>. Consequently, we formalize the detection of CP as a binary classification task. The input for the task includes various user signals, i.e., sentiment judgment, behavior responses, and brain signals. The target is the bias assigned to each user-personage pair, i.e., whether the presented videos have a positive or negative polarity.

During the field study, we designed a setting to reflect positive or negative polarization for each participant and personage pair. For example, when Participant A and personage Michael Corleone were in the setting of positive opinion polarization, participant A was consistently shown videos positively depicting Mike. For this specific pair, participant A's EEG and behavioral responses were recorded while watching Mike-themed videos in both pre-study and post-study phases. Additionally, sentiment scores regarding Mike, collected at two different points in the sentiment collection phase, were incorporated. The assigned label for this data pair is "1," indicating a positive opinion polarization. In contrast, a label of "0" denotes negative opinion polarization.

### *Feature extraction*

During the pre-study and post-study, continuous collection of Electroencephalogram (EEG) physiological signals was conducted on each participant. In this study, we utilized a 64-channel Quik-Cap (Compumedics NeuroScan) for EEG signal acquisition, with electrode placement following the International 10–20 system<sup>55</sup>. The collected EEG data underwent preprocessing for further analysis, which included re-referencing to the average mastoid, baseline correction, low-pass filtering at 50 Hz, high-pass filtering at 0.5 Hz, and artifact removal. For artifact removal, we implemented a parametric noise covariance model described in<sup>56</sup>, which models the spatiotemporal noise covariance as a Kronecker product of spatial and temporal matrices. This approach effectively characterizes various noise sources including ocular, cardiac, and muscular artifacts while preserving the neural signals of interest.

To utilize EEG data as features for training our model, we extract Differential Entropy (DE)<sup>57</sup> as input feature. DE is a crucial tool in assessing EEG signals. DE was calculated as follows:  $DE = - \int P(f) \log(P(f)) df$ , where  $f$  represents frequency ranging from 1 to 50, and  $P(f)$  denotes the power spectral density (PSD) of EEG signals. This calculation of PSD was conducted using Welch's method<sup>58</sup>, which employs a moving window to extract spectral density. The length of this window was set to twice the inverse of the frequency band's lower limit, with the system operating at a sampling rate of 1000 Hz.

### *Feature selection*

We have implemented three input user signals and their combinations as distinct input features: (1) Behavior response Collected through our platform, including "likes" and "viewing duration ratio". (2) EEG features in terms of Differential Entropy (DE) collected at every electrode across five passbands, following the method described in Method Feature Extraction. (3) Participants' sentiment toward the personages expressed as scores on a five-point scale (1–5), collected with two questionnaires. In the classification task, EEG feature selection was essential due to the substantially larger dimensions of EEG features compared to others. For each user-personage pair, we included EEG features collected during four different periods. The time periods are participants viewing a positive/negative video corresponding to a specific personage (denoted as positive videos and negative videos) in the pre-study and post-study, respectively. The dimensionality of each EEG feature was  $30 * 62 * 5$  (i.e., #time windows \* #electrodes \* #bands). To leverage the most effective EEG features, we selected the alpha frequency band, which is closely associated with emotions and cognitive functions<sup>59</sup>. For prediction purposes, the FPz electrode in the alpha band demonstrated optimal performance in our cross-validation tests, despite its absence from significant electrodes in correlation analysis. This difference stems from the prediction model's use of both Lab Studies' data and non-linear modeling (validation results in SI Appendix Figure 10, Table 7). When restricting prediction to only Lab Study 2 data (matching our correlation analysis conditions), electrodes showing significant correlations (C5, FC5) performed better than FPz, confirming the methodological basis for these differences (SI Appendix Table 8). Based on both prior literature supporting the role of FPz and its superior performance in our cross-validation tests, we selected the FPz alpha-band feature for our primary prediction model. We selected the "FPz" feature in the alpha band guided by Dixon et al. (2017)<sup>60</sup>, who emphasized its relevance to emotional processing and focus. It is important to note that this selection of the FPz electrode specifically applies to our prediction model, not to the correlation analysis presented in the Results section, where all electrodes were examined for significant effects. Considering the difficulties in collecting and processing multi-channel signals

for large-scale commercial devices and practical scenarios, our work focuses on CP predictions based on single-channel signals. To represent user-personage level information, we averaged each feature over time (i.e., average (30,62,5) to (62,5)), followed by the selection of specific frequency bands and electrodes.

#### Classification model

In our research, we explored various machine models before ultimately selecting the XGBoost model<sup>61</sup>, recognized for its unique characteristics and benefits. We used the default parameter settings of XGBoost based on the Python scikit-learn library. XGBoost, an acronym for eXtreme Gradient Boosting, represents an advanced application of gradient boosting algorithms. XGBoost has shown its strong performance and stability across diverse datasets. Notably, XGBoost excels in efficiency and effectiveness, aligning seamlessly with the specific requirements of our study. The high cost and time constraints associated with EEG data collection resulted in a relatively limited set of features. Consequently, we refrained from employing more complex models. Future research could potentially expand the dataset and explore the application of more sophisticated models.

#### Data availability

Raw data for the study is available at: <https://figshare.com/s/2312a436f38fdd8373e2>

#### Code availability

Code is available at: [https://anonymous.4open.science/r/EEG\\_and\\_Explicit\\_Feedback\\_for\\_Opinion\\_Polarization\\_Detection-0E18](https://anonymous.4open.science/r/EEG_and_Explicit_Feedback_for_Opinion_Polarization_Detection-0E18)

Received: 11 December 2024; Accepted: 13 May 2025

Published online: 25 May 2025

#### References

- Jamieson, K. H. *Packaging the presidency: A history and criticism of presidential campaign advertising* (Oxford University Press, 1996).
- Macy, M. W., Ma, M., Tabin, D. R., Gao, J. & Szymanski, B. K. Polarization and tipping points. *Proceedings of the National Academy of Sciences* **118**, e2102144118 (2021).
- Holone, H. The filter bubble and its effect on online personal health information. *Croatian medical journal* **57**, 298 (2016).
- McCright, A. M. & Dunlap, R. E. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly* **52**, 155–194 (2011).
- Levy, R. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review* **111**, 831–870 (2021).
- Arora, S. D., Singh, G. P., Chakraborty, A. & Maity, M. Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change* **183**, 121942 (2022).
- Alatawi, F. et al. A survey on echo chambers on social media: Description, detection and mitigation. arXiv preprint [arXiv:2112.05084](https://arxiv.org/abs/2112.05084) (2021).
- Pariser, E. *The filter bubble: What the Internet is hiding from you* (penguin UK, 2011).
- Jamieson, K. H. & Cappella, J. N. *Echo chamber: Rush Limbaugh and the conservative media establishment* (Oxford University Press, 2008).
- Valentino, N. A., Banks, A. J., Hutchings, V. L. & Davis, A. K. Selective exposure in the internet age: The interaction between anxiety and information utility. *Political Psychology* **30**, 591–613 (2009).
- Sunstein, C. R. *Infotopia: How Many Minds Produce Knowledge* (Oxford University Press, 2006).
- Arguedas, R., Robertson, C., Fletcher, R. & Nielsen, R. Echo chambers, filter bubbles, and polarisation: a literature review. *Reuters Institute for the Study of Journalism* (2022).
- Dahlgren, P. M. A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review* **42**, 15–33. <https://doi.org/10.2478/nor-2021-0002> (2021).
- Eady, G., Nagler, J., Guess, A., Zilinsky, J. & Tucker, J. A. How many people live in political bubbles on social media? evidence from linked survey and twitter data. *Sage Open* **9**, <https://doi.org/10.1177/2158244019832705> (2019).
- Li, N. et al. An exploratory study of information cocoon on short-form video platform. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, 4178–4182 (Association for Computing Machinery, New York, NY, USA, 2022).
- Piao, J., Liu, J., Zhang, F., Su, J. & Li, Y. Human-ai adaptive dynamics drives the emergence of information cocoons. *Nature Machine Intelligence* **5**, 1214–1224. <https://doi.org/10.1038/s42256-023-00731-4> (2023).
- Waller, I. & Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature* **600**, 264–268. <https://doi.org/10.1038/s41586-021-04167-x> (2021).
- Xing, Y., Wang, X., Qiu, C., Li, Y. & He, W. Research on opinion polarization by big data analytics capabilities in online social networks. *Technology in Society* **68**, 101902. <https://doi.org/10.1016/j.techsoc.2022.101902> (2022).
- Chen, M. Exploration in recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 551–553 (2021).
- Chen, M. et al. Values of user exploration in recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 85–95 (2021).
- Einav, G., Allen, O., Gur, T., Maaravi, Y. & Ravner, D. Bursting filter bubbles in a digital age: Opening minds and reducing opinion polarization through digital platforms. *Technology in Society* **71**, 102136 (2022).
- Combs, A. et al. Reducing political polarization in the United States with a mobile chat platform. *Nature human behaviour* **7**, 1454–1461 (2023).
- Strandberg, K., Himmelroos, S. & Grönlund, K. Do discussions in like-minded groups necessarily lead to more extreme opinions? deliberative democracy and group polarization. *International Political Science Review* **40**, 41–57 (2019).
- Criss, S., Michaels, E. K., Solomon, K., Allen, A. M. & Nguyen, T. T. Twitter fingers and echo chambers: Exploring expressions and experiences of online racism using twitter. *Journal of Racial and Ethnic Health Disparities* **8**, 1322–1331 (2021).
- Mäs, M. & Flache, A. Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PloS one* **8**, e74516 (2013).
- Lee, F. L. Impact of social media on opinion polarization in varying times. *Communication and the Public* **1**, 56–71 (2016).

27. Peng, A., Liu, J. & Gao, Q. Public opinion analysis strategy of short video content review in big data environment. In *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 100–104, <http://doi.org/10.1109/ICCWAMTIP47768.2019.9067687> (Chengdu, China, 2019).
28. Ye, Z. *et al.* Relevance feedback with brain signals. *ACM Transactions on Information Systems* (2023).
29. Moshfeghi, Y., Pinto, L. R., Pollick, F. E. & Spampinato, J. M. Understanding relevance: An fmri study. In *European conference on information retrieval*, 14–25 (Springer, 2013).
30. Moshfeghi, Y., Triantafyllou, P. & Pollick, F. Towards predicting a realisation of an information need based on brain signals. In *The World Wide Web Conference*, 1300–1309 (2019).
31. Pinkosova, Z., McGeown, W. J. & Moshfeghi, Y. Moderating effects of self-perceived knowledge in a relevance assessment task: An eeg study. *Computers in Human Behavior Reports* **11**, 100295 (2023).
32. Dirican, A. C. & Göktürk, M. Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science* **3**, 1361–1367 (2011).
33. He, Z. *et al.* Understanding user immersion in online short video interaction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 731–740 (2023).
34. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia* **126**, 1763–1768 (2018).
35. Maris, E. & Oostenveld, R. Nonparametric statistical testing of eeg- and meg-data. *Journal of neuroscience methods* **164**, 177–190 (2007).
36. Ertl, M., Hildebrandt, M., Ourina, K., Leicht, G. & Mulert, C. Emotion regulation by cognitive reappraisal—the role of frontal theta oscillations. *NeuroImage* **81**, 412–421 (2013).
37. Buzsáki, G. & Wang, X.-J. Mechanisms of gamma oscillations. *Annual review of neuroscience* **35**, 203–225 (2012).
38. Yang, K. *et al.* High gamma band eeg closely related to emotion: evidence from functional network. *Frontiers in human neuroscience* **14**, 89 (2020).
39. Huang, J. & Ling, C. X. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* **17**, 299–310 (2005).
40. Magnus, O. & Van der Holst, M. Zeta waves: a special type of slow delta waves. *Electroencephalography and clinical neurophysiology* **67**, 140–146 (1987).
41. Knyazev, G. G. Eeg delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience & Biobehavioral Reviews* **36**, 677–695 (2012).
42. Knyazev, G. G. Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neuroscience & Biobehavioral Reviews* **31**, 377–395 (2007).
43. Harmony, T. The functional significance of delta oscillations in cognitive processing. *Frontiers in integrative neuroscience* **7**, 83 (2013).
44. Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I. & Schroeder, C. E. Entrainment of neuronal oscillations as a mechanism of attentional selection. *science* **320**, 110–113 (2008).
45. Aftanas, L. I. & Golocheikine, S. A. Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution eeg investigation of meditation. *Neuroscience letters* **310**, 57–60 (2001).
46. Mölle, M. & Born, J. Slow oscillations orchestrating fast oscillations and memory consolidation. *Progress in brain research* **193**, 93–110 (2011).
47. Headley, D. B. & Weinberger, N. M. Gamma-band activation predicts both associative memory and cortical plasticity. *Journal of Neuroscience* **31**, 12748–12758 (2011).
48. Kumar, G. P., Sharma, K., Ramakrishnan, A. & Adarsh, A. Increased entropy of gamma oscillations in the frontal region during meditation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 787–790 (IEEE, 2021).
49. Erk, S. *et al.* Acute and sustained effects of cognitive emotion regulation in major depression. *Journal of Neuroscience* **30**, 15726–15734 (2010).
50. Phelps, E. A. & LeDoux, J. E. Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron* **48**, 175–187 (2005).
51. Headley, D. B. & Paré, D. In sync: gamma oscillations and emotional memory. *Frontiers in behavioral neuroscience* **7**, 170 (2013).
52. Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L. & Cichocki, A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics* **49**, 1110–1122 (2018).
53. Zawacki, E. E., Bohon, W., Johnson, S. & Charlevoix, D. J. Exploring tiktok as a promising platform for geoscience communication. *Geoscience Communication* **5**, 363–380 (2022).
54. Li, Y. & Zhao, H. Evaluation and control of opinion polarization and disagreement: a review. *International Journal of Crowd Science* **5**, 114–128 (2021).
55. Homan, R. W., Herman, J. & Purdy, P. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology* **66**, 376–382 (1987).
56. Huizenga, H. M., De Munck, J. C., Waldorp, L. J. & Grasman, R. P. Spatiotemporal eeg/meg source analysis based on a parametric noise covariance model. *IEEE Transactions on Biomedical Engineering* **49**, 533–539 (2002).
57. Duan, R.-N., Zhu, J.-Y. & Lu, B.-L. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 81–84 (IEEE, 2013).
58. Welch, P. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics* **15**, 70–73 (1967).
59. Klimesch, W. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews* **29**, 169–195 (1999).
60. Dixon, M. L., Thiruchselvam, R., Todd, R. & Christoff, K. Emotion and the prefrontal cortex: An integrative review. *Psychological bulletin* **143**, 1033 (2017).
61. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

## Acknowledgements

We would like to express our gratitude to Shaorun Zhang and Zhiyu He for their assistance with using the experimental equipment and data analysis. We also appreciate the encouragement and help from other professors and colleagues at THUIR lab.

## Author contributions

The study was designed by Bangde Du, Ziyi Ye, Zhijing Wu, Qingyao Ai and Yiqun Liu; execution carried out by Bangde Du; analysis by Bangde Du, Ziyi Ye, and Monika Jankowska; manuscript writing by Bangde Du, Ziyi Ye, Monika Jankowska, Zhijing Wu, Qingyao Ai, and Yujia Zhou; experimental resources provided by Yiqun Liu.

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-02423-y>.

**Correspondence** and requests for materials should be addressed to Q.A. or Y.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025