iNews: A Multimodal Dataset for Modeling Personalized Affective Responses to News

Anonymous ACL submission

Abstract

Current approaches to emotion detection often overlook the inherent subjectivity of affective experiences, instead relying on aggregated labels that mask individual variations in emotional responses. We introduce iNews, a novel large-scale dataset explicitly capturing subjective affective responses to news headlines. Our dataset comprises annotations from 291 demographically diverse UK participants across 2,899 multimodal Facebook news posts from major UK outlets, with an average of 5.18 annotators per sample. For each post, annotators provide multifaceted labels including valence, arousal, dominance, discrete emotions, content relevance judgments, sharing likelihood, and modality importance ratings (text, image, or both). Furthermore, we collect comprehensive annotator persona information covering demographics, personality, media trust, and consumption patterns, which explain 15.2% of annotation variance - higher than existing NLP datasets. Incorporating this information yields a 7% accuracy gain in zero-shot prediction and remains beneficial even with 32-shot. iNews will enhance research in LLM personalization, subjectivity, affective computing, and individual-level behavior simulation[1].

1 Introduction

Emotion is a quintessential example of the human subjectivity that permeates human experience, profoundly shaping our perceptions, decisions, and interactions (Zadra and Clore, 2011; Lerner et al., 2015, *inter alia*). As large language models (LLMs) achieve widespread global adoption, understanding and modeling these subjective affective responses becomes crucial for developing systems that can effectively personalize interactions and ensure alignment with diverse human values and preferences. Affective responses vary significantly across individuals, shaped by a complex interplay of individual and group-level factors, including age, gender, personality, cultural backgrounds, and lived experiences (Kring and Gordon, 1998; Costa and McCrae, 2008; Charles and Carstensen, 2010; Mesquita and Frijda, 1992).

Despite the well-established understanding in psychology of the inherent subjectivity of emotion, many natural language processing (NLP) approaches to emotion detection fall short of embracing a user-centric perspective that accounts for individual differences. They often rely on aggregation techniques, such as majority voting or averaging, which obscure nuanced individual reactions and neglect to explicitly define the intended perspective of the analysis (e.g., writer vs. reader) (Ovesdotter Alm, 2011; Plank, 2022; Cabitza et al., 2023). Such aggregation, without acknowledging individual variability and potential biases, risks generating misleading conclusions and hindering the development of personalized and user-centric systems.

To address these limitations, we introduce iNews, a novel large-scale dataset specifically designed to capture the inherent subjectivity of affective responses to real-world news content (overview in Figure 1). Our dataset comprises fine-grained affective responses from 291 annotators to 2,899 Facebook posts from leading UK media outlets. The annotations include: valence-arousal-dominance (VAD) ratings (Mehrabian and Russell, 1974; Bradley and Lang, 1994), Plutchik's eight basic emotions (Plutchik, 1980), perceived post relevance, modality importance, and sharing likelihood. In addition, we collect a comprehensive set of annotator characteristics[2] (e.g. demographics, personality, media consumption habits), drawing upon insights from the differential media effects literature.

[1]The dataset will be made publicly available at [HuggingFace Link].

[2]Throughout this work, we also use the term persona information to refer to the same concept.

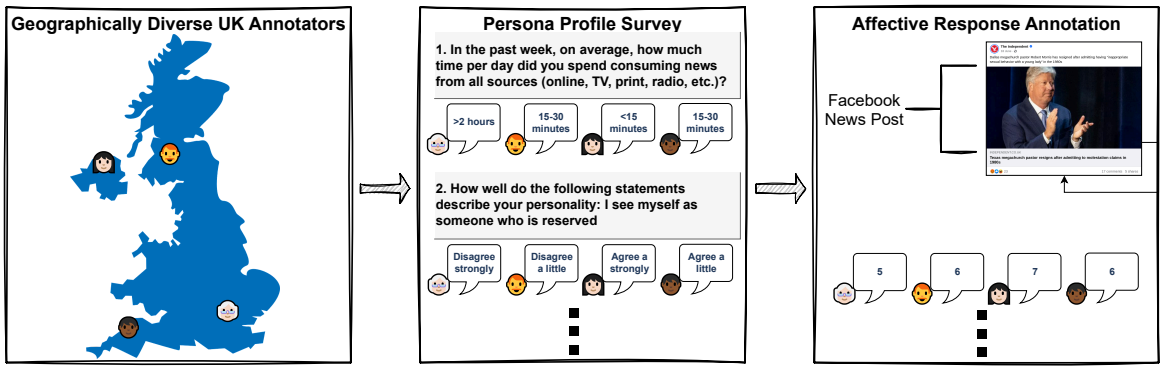


Figure 1: Overview of the data collection process. The process involves three main stages: (1) we recruit demographically diverse UK annotators; (2) annotators complete a persona profile survey capturing demographics, ideology, news consumption, cognitive traits, personality, and emotional characteristics; and (3) annotators provide affective response annotations for Facebook news posts, including valence, arousal, dominance, discrete emotions, modality influence, personal relevance, and sharing likelihood.

Our regression analysis confirms that annotator characteristics explain a substantial portion of the annotation variance (15.2% - higher than observed in any NLP dataset to date), highlighting the importance of incorporating individual differences when modeling subjective phenomena like affect. Furthermore, through an open-ended questionnaire with a subset of annotators ($N = 20$), we identify nuanced patterns in how individuals experience and articulate their emotional reactions to news content, beyond the scope of our structured annotations and survey.

In a case study demonstrating the practical value of this rich persona information, we show that incorporating annotator characteristics can improve LLM predictions of individual-level affective responses by up to 7% in accuracy in zero-shot settings, although overall accuracy remains relatively modest (around 40%). When comparing different input modalities (image vs. text), we find that image inputs typically outperform text in zero-shot scenarios but this advantage diminishes in few-shot settings. In the few-shot setting, we observe the "early descent phenomenon" (Lin and Lee, 2024; Agarwal et al., 2024), where performance initially dips below zero-shot levels with very few examples before improving as the number of shots increases. We ultimately reach 44.4% accuracy at 32-shot. Even at this level, incorporating persona information still yields a further performance gain, suggesting that persona-based and example-based approaches provide complementary signals for modeling individual differences.

The iNews dataset stands to benefit a wide range of research areas: for affective computing researchers modeling emotion recognition while account for individual differences; for LLM developers advancing personalization and subjective phenomena handling; for human behavior simulation researchers modeling individual-level information processing; for social computing scholars investigating demographic effects in content presentation; and for AI alignment researchers studying preference diversity across human populations.

2 Related Work

2.1 News, Emotion, and Individual Differences

The interplay between news content, emotional responses, and downstream cognitive and behavioral effects is often an area of focus in communication and psychology. Prior research establishes that news often exhibits a negativity bias, eliciting negative emotions and heightened arousal in readers (Soroka et al., 2019). However, individual responses vary considerably based on demographic factors, pre-existing political attitudes and identities, personality traits, and other individual and group-level characteristics (Oliver, 2002; Valkenburg and Peter, 2013; Soroka et al., 2019). This heterogeneity carries significance beyond immediate emotional experiences, fundamentally influencing information processing and behavior. Emotions provide evaluative feedback, impacting veracity judgments (Martel et al., 2020) and shaping reasoning and decision-making (Marcus et al., 2000; Storbeck and Clore, 2008).

Existing research on the affective dimension of news perception predominantly focuses on the emotional tone of the news content itself, rather than the induced emotional responses of individual readers (de Hoog and Verboon, 2020). Much of this work relies on aggregate-level analysis, obscuring individual-level variation. Our work addresses these limitations by redirecting attention to fine-grained reader responses. We present a large-scale dataset designed to capture and analyze the spectrum of individual affective responses to news headlines, facilitating a more nuanced understanding of the relationship between news, emotion, and individual differences.

2.2 Emotion Detection in NLP

Emotion detection has been a long-standing focus within NLP (Strapparava and Mihalcea, 2007; Plaza-del Arco et al., 2024). Recent years have seen a large number of valuable resources on the task (see Demszky et al. (2020); Oberländer et al. (2020); Plaza del Arco et al. (2020) for a overview). These efforts have significantly advanced the field, leading to more accurate and robust emotion detection systems.

However, most existing datasets rely on aggregated "gold labels", overlooking the inherent subjectivity and variation in human emotional perception (Ovesdotter Alm, 2011; Plank, 2022; Cabitza et al., 2023). Extensive psychological research demonstrates the significant influence of both individual characteristics (e.g., age, gender, personality traits) and group-level factors (e.g., cultural background) on how we perceive and interpret emotions (Mesquita and Frijda, 1992; Kring and Gordon, 1998; Costa and McCrae, 2008; Charles and Carstensen, 2010). Consequently, models trained on datasets with aggregated labels inevitably fail to capture the nuanced, individualized nature of affective responses. This limits their effectiveness in real-world applications that demand personalized understanding and responsiveness to diverse emotional expressions.

Limited attempts have been made to incorporate annotator background information (Plaza-del Arco et al., 2024). For instance, Diaz et al. (2018) provide demographic data alongside sentiment annotations in an online community dataset; however, this work is limited by its focus on sentiment (rather than fine-grained emotions), its restriction to a specific online community, and its lack of multi-faceted affective response measures. To our knowledge, no existing dataset combines comprehensive individual difference variables, fine-grained affective responses, and annotations of real-world news content, as ours does.

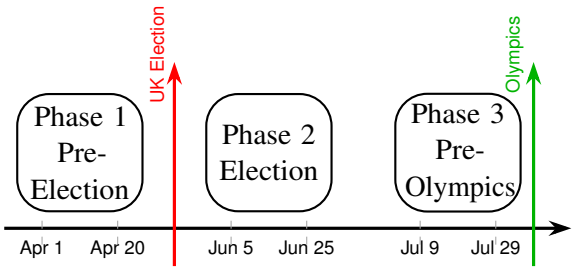


Figure 2: Data collection timeline, with the 2024 UK General Election and 2024 Paris Olympics marked out.

3 Dataset Collection Protocol

To address the limitations of existing emotion detection datasets, we develop a two-stage data collection protocol to capture individualized affective responses to news headlines (Figure 1). Our protocol emphasizes ecological validity and the collection of rich persona variables to enable the study of how personal characteristics influence affective responses.

Sampling Our dataset comprises annotations of Facebook news posts, collected in three phases to capture diverse news contexts surrounding the 2024 UK general election and the Paris Olympics (see Figure 2). These phases are: Phase 1 (April 1-20), the pre-election period; Phase 2 (June 5-25), the election campaign period after Parliament's dissolution; and Phase 3 (July 9-29), the post-election and pre-Olympics period.

This three-phase design ensures temporal diversity and mitigates the influence of any single major event on our findings. We initially used random sampling (Phase 1) to gather a broad sample. Recognizing that some outlets are far more prolific but have lower engagement, we transitioned to stratified sampling (Phases 2 and 3) proportional to each outlet's follower count. This approach maximizes ecological validity by ensuring our sample reflects the news content that readers are actually likely to encounter.

For each phase, we collect news posts via CrowdTangle. While social media content may not represent an outlet's entire output, we posit that these posts reflect editorial choices and the outlet's intended public image. Each post typically includes

an image, a short description, and the headline, with the image linking to the full article (see Figure 12 for an example). To ensure maximal ecological validity and minimize bias, we present screenshots instead of text of the posts to annotators, capturing reaction counts but excluding comments to avoid influencing annotator responses. The decision to use screenshots rather than just headlines is also supported by a pilot study (Section A.2) demonstrating significant differences in affective responses based on presentation modality.

Annotator Recruitment We recruit annotators through Prolific, using quota sampling to ensure a relatively balanced representation across gender, age, political leaning, and UK geographical regions (see Figure 6 and Table 3 for details). Each of the 291 annotators contribute annotations for approximately 50 headlines. The annotation process takes around 45 minutes. Annotators are compensated £8.58, in accordance with the UK National Living Wage at the time of the data collection.

Stage 1: Persona Profile Survey This stage, implemented in Qualtrics, gathers background information ("persona variables") about each annotator. The survey incorporates validated items from well-established questionnaires (Ofcom, 2024a; Reuters Institute, 2024), alongside standard psychological instruments. These variables (detailed in Table 3) are selected to capture individual differences known to influence news interpretation and emotional responses, enabling us to study how these factors mediate affective reactions.

The collected variables span five key areas: **Demographics and Ideology** captures age, gender, and political ideology. **News Consumption and Trust** measures consumption patterns and trust ratings for major UK news outlets. **Cognitive Traits** are assessed through the Cognitive Reflection Test (CRT) (Frederick, 2005). **Personality Traits** are measured using the 10-item Big Five Inventory (BFI-10) (Rammstedt and John, 2007). **Emotional Characteristics** are evaluated using both the Perth Emotional Reactivity Scale (PERS) (Preece et al., 2018) and the Positive and Negative Affect Schedule (PANAS) (Crawford and Henry, 2004).

Stage 2: Headline Annotation Annotators are provided with detailed guidelines, adapted from Bradley and Lang (2007), which are accessible throughout the annotation process[3]. The annotation interface is built using the Potato annotation tool (Pei et al., 2022).

We then present annotators with news posts public Facebook pages of major UK outlets (see Table 5). For each news post screenshot (see Figure 8 for exact question wording), annotators provide five types of responses: **Dimensional Emotion Ratings** capture valence, arousal, and dominance on a Likert scale of 1-7 using the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994). **Discrete Emotion Classifications** involve categorizing into one of Plutchik's eight basic emotion (Plutchik, 1980). **Modality Influence** assesses the relative influence of the image versus the text on their emotional response. **Personal Relevance** rates the headline's personal relevance. **Sharing Likelihood** measures the likelihood they would share the post.

4 Descriptive Analysis

Our dataset comprises 2,899 annotated news posts, with an average of 5.18 annotations per post from 291 distinct annotators. Each annotator also provides ratings for three standardized items from the Affective Norms for English Text (ANET) (Bradley and Lang, 2007).

Annotator Demographics Our annotator pool exhibits diversity across gender, political ideology, ethnicity, education levels, and other key demographic variables. Crucially, we have annotators from 97 out of 124 UK postcode areas, ensuring substantial geographic diversity within the UK. See Table 3 for a comprehensive breakdown of annotator characteristics.

Distribution of Annotations Figure 7 presents the distributions of the collected annotation variables. Key observations include: The neutral value (4) is the most frequent for all three dimensions. As expected, the valence scores tend to skew negatively, arousal scores are predominantly high, and dominance scores skew slightly low. For discrete emotions, "neutral" is the most commonly selected emotion, followed by "sad". Interestingly, the next most frequent emotion is "happy," which is likely due to the limitation of having only one category for positive emotions. Further details, including distributions for relevance, sharing likelihood,

[3]Link: https://docs.google.com/document/d/1RPkjaPSksRbCy3y5d4WltidcUGhlH_np-aAuY2eH33c/

and modality influence, are available in Appendix Section A.4.1. Additionally, we analyze inter-annotator agreement in Section A.5, finding Krippendorff's α values comparable to existing emotion annotation datasets, with moderate agreement for valence ($\alpha = 0.468$) and lower agreement for arousal ($\alpha = 0.145$) and dominance ($\alpha = 0.203$).

Outlet-level Analysis We present the summary statistics of affect annotations across news outlets in Table 5. All outlets are on average more negative content (low valence; with discrete emotions predominantly categorized as either neutral or sad/angry) while maintaining higher-than-neutral levels of arousal. See Section A.4.2 for a comparison between broadsheet and tabloid outlets.

News Post Characteristics To analyze the topical composition of news posts, we employ the IPTC NewsCode taxonomy (International Press Telecommunications Council, 2024), a widely-adopted industry standard for news categorization. We choose this established taxonomy over topic modeling given the well-defined nature of news categorization as a task. We classify news post using zero-shot with Gemini 1.5 Pro (prompt in Section A.6). Figure 10 shows the topic distribution, and Figure 11 shows mean arousal per topic. The most frequent categories are arts/culture/entertainment/-media (25.4%), crime/law/justice (12.9%), and politics (9.6%). This prevalence of hard news over soft news aligns with prior research on media organizations' social media strategies (Lamot, 2022) and platform-specific characteristics of Facebook (Newman et al., 2015). As expected, arousal is higher for topics like conflict/war (4.83) and disasters/accidents (4.77) compared to arts/culture (3.85), consistent with previous findings (Soroka et al., 2019).

5 Regression Analysis

To quantify the influence of individual differences on affective responses, and to assess the effectiveness of our collected persona variables in capturing these differences, we conduct a regression analysis using linear mixed-effects models, focusing on the arousal dimension as a case study (Likert scale, 1-7).

Models We construct three models to systematically decompose the variance in affective responses: (1) a **Null Model** with only news text as a random effect, serving as our baseline; (2) a **Persona Model** adding 47 persona variables as fixed effects while controlling for text effects; and (3) a **User Model** incorporating both news text and user ID as random effects to capture all user-level variance, including unobserved individual differences.

We evaluate each model using both marginal R^2 (variance explained by fixed effects) and conditional R^2 (variance explained by fixed and random effects) show the results in Table 1.

Strong explanatory power of persona variables. News content alone explains 13.1% of the variance in arousal ratings (null model, conditional $R^2 = 0.131$). Incorporating our collected persona variables significantly increases the explained variance to 28.6%. This improvement, higher than that observed in existing NLP datasets with annotator characteristics (Diaz et al., 2018; Hu and Collier, 2024), underscores the importance of individual differences in modeling subjective phenomena and validates the richness of the persona information collected in iNews.

Unobserved individual factors still matters. Despite explicitly modeling a comprehensive set of persona variables, the User model explains more variance than the Persona model (0.317 vs. 0.286). This gap suggests the presence of additional unobserved individual factors that modulate affective responses—factors that remain unaccounted for even with our extensive variable collection.

Persona information matter in modeling affective responses. Our findings demonstrate that modeling individual differences is crucial for understanding affective responses to text. The persona variables collected in our iNews dataset capture a large portion of this individual variability, validating our data collection protocol and demonstrating the dataset's value for advancing personalized language technologies. The remaining unexplained variance highlights both the inherent complexity of human affect and the potential for future research to contextualize additional contributing factors.

6 Qualitative Analysis of Post-Annotation Questionnaire

To complement our quantitative analysis of persona variables and gain a richer understanding of how individual differences shape emotional responses, we conduct a post-annotation qualitative study. Twenty annotators from our main study complete an open-ended questionnaire (administered

Model	Fixed Effects	Random Effects	Marginal R^2	Conditional R^2
Null	None	Text	0.000	0.131
Persona	Persona variables (47)	Text	0.152	0.286
User	None	Text + User	0.000	0.317

Table 1: Comparison of mixed-effects regression models for predicting affective arousal ratings. Marginal R^2 indicates variance explained by fixed effects alone, while Conditional R^2 shows the total variance explained by both fixed and random effects.

via Qualtrics/Prolific), consisting of six open-ended questions probing how readers process and respond to news content (see Section A.8 for questions and expanded analysis).

We perform a thematic analysis of the responses, employing a systematic coding approach facilitated by an LLM. The analysis reveals insights that help contextualize the individual differences observed in our survey data. For instance, one annotator describe the influence of growing up during the cold war on their emotional responses. Another highlights how their working-class background leads them to be "kind of numb to some types of news," while still emphasizing the emotional impact of "people getting hurt for no reason." The news platform itself emerges as a mediating factor, with one participant stating, "I don't really buy what I see on Facebook, so it doesn't get to me as much." These rich self-narratives, combined with structured persona data, could inform more nuanced models of individual differences in news response, aligning with recent work on using qualitative interview data to simulate human behavior (Park et al., 2024).

7 Predicting Individual Affective Arousal

Building on our regression analysis, we now investigate the capacity of current LLMs to predict individual-level affective response. We continue to focus on the emotional arousal dimension as a case study, examining how well models estimate specific annotators' responses under various zero-shot and few-shot conditions.

7.1 Experimental Setup

We randomly sample 30 annotators from iNews dataset. For each annotator, we reserve 32 of their annotated posts for potential few-shot demonstrations and utilize the remaining posts (579 samples total) for testing. For evaluation, we employ three complementary metrics that capture different aspects of prediction quality: Mean Absolute Error (MAE) to measure overall prediction accuracy, Exact Accuracy to identify precise matches with annotator ratings, and ±1 Accuracy (the percentage of predictions falling within one point of the ground truth) to account for the inherent subjectivity in emotional assessment by allowing slight variations. Our evaluation compares model predictions against each individual annotator's ratings.

We conduct experiments across seven frontier models, including both API-based models [Gemini 1.5 Pro (Team et al., 2024), GPT-4o (Hurst et al., 2024), Grok-2 (xAI, 2025)] and open-weight models [Llama-3.2-90B-Vision (Meta AI, 2024a), Qwen2.5-VL-72B-Instruct (QwenLM, 2025), Llama-3.3-70B-Instruct (Meta AI, 2024b), Llama-3.1-405B-Instruct (Meta AI, 2024c)], with all except the last two capable of processing multimodal inputs. In rare cases where formatting or safety concerns prevents a model from generating a prediction, we assign -1 as the prediction to penalize such behavior. As we decode only a single token for the answer, temperature settings and sampling parameters are not relevant to this process.

We examine four input conditions. The text-only condition provides a detailed textual description of each news post, while the image-only condition uses the original news screenshot. We then augment each of these base conditions with persona information, creating text-with-persona and image-with-persona conditions where annotator characteristics are incorporated into the system prompt. Since our annotators originally rated news screenshots, we leverage Gemini 1.5 Pro to generate comprehensive textual descriptions for the text-only conditions, enriching these with headline text and engagement metrics. The complete prompt templates and an illustrative news post image-text pair are provided in Sections A.10 and A.9, respectively. We present our zero-shot evaluation results in Table 2.

While persona variables provide valuable signals for personalization, they inevitably offer an incomplete view of individual preferences and behaviors, as the richness and complexity of human experience extends far beyond what can be cap-

tured through demographic and personality questionnaires (Dong et al., 2024). We hypothesize that incorporating behavioral data, specifically, an individual's prior annotations, could provide complementary information for modeling affective responses. To test this hypothesis, we conduct k-shot experiments ($k \in \{4, 8, 16, 32\}$) with and without persona information across both text and image modalities. Figure 3 presents the Exact Accuracy results for Gemini 1.5 Pro, our top-performing zero-shot model (complete results in Table 9 and Figure 13). Due to resource constraints, we are only able to conduct few-shot experiments with Gemini 1.5 Pro.

7.2 Zero-shot Evaluation

Current LLMs demonstrate reasonable default zero-shot alignment with UK annotators. Without personalization, models achieve seemingly encouraging baseline performance with ±1 accuracy exceeding 70%. However, this metric alone can be misleading - a naive predictor that simply outputs the population mean would likely achieve similar ±1 accuracy given the roughly Gaussian distribution of arousal ratings (see Figure 7). The consistently low exact accuracy (< 40%) across all models provides a more stringent evaluation of true personalization capability. This substantial gap between ±1 and exact accuracy suggests that while models can broadly approximate the range of typical responses, they struggle to capture individual-specific variations in emotional reactions.

Incorporating persona information consistently improves performance. The improvements are particularly large for Gemini 1.5 Pro, where persona information reduces MAE by 11.6% (1.034 → 0.914) for text input and 10.1% (0.936 → 0.841) for image input. These substantial gains demonstrate that current LLMs can effectively leverage explicit persona variables to better simulate individual annotators, validating our persona variable collection strategy. This result aligns with prior work on the effectiveness of persona prompting (Rescala et al., 2024; Dong et al., 2024; Hu and Collier, 2024).

Image inputs consistently outperform textual inputs. Our analysis reveals a clear advantage for image-based prediction across all vision-language models except Llama-3.2-90B-Vision. The optimal performance is achieved by Gemini 1.5 Pro with image input and persona information (MAE: 0.841, ±1 Accuracy: 82.04%), surpassing the best text-only configuration from Llama-3.1-405B-Instruct (MAE: 0.885, ±1 Accuracy: 81.17%). This superiority of visual inputs aligns with prior working documenting stronger affective responses to images versus text in psychology and communication literature (Iyer and Oldmeadow, 2006; Powell et al., 2015). Even our high-quality textual descriptions (example in Section A.9), appear unable to fully capture the affective richness encoded in visual stimuli. This observation is echoed by annotators who report particularly intense emotional reactions to images of suffering or tragedy (see Section A.8.2).

Models exhibit varying degrees of steerability through persona prompting. Gemini 1.5 Pro, Grok-2, Qwen2.5-VL-72B-Instruct and the Llama family show high responsiveness to persona information, while GPT-4o maintains more consistent behavior with and without persona information. This variation suggests fundamental differences in these models' capacity for steerably pluralistic alignment (Sorensen et al., 2024).

7.3 Few-shot Evaluation

Few-shot learning demonstrates a consistent pattern of initial performance degradation followed by gradual recovery. For both text and image modalities, we observe performance drop when transitioning from zero-shot to 4-shot setting. Performance gradually recovers with increasing demonstrations, with text-input models surpassing zero-shot performance at 8 shots (no persona) or 16 shots (with persona), continuing to improve up to 32 shots. However, the image modality shows a sharper initial decline and slower recovery, only matching zero-shot performance at 32 shots for exact accuracy and still lagging in ±1 Accuracy and MAE. This initial deterioration aligns with the early ascent (in terms of risk) phenomenon in in-context learning (Lin and Lee, 2024; Agarwal et al., 2024), where models initially struggle to effectively integrate limited demonstrations. We hypothesize that the inherent subjectivity and noise in emotional arousal annotations may exacerbate this effect, leading to overfitting with sparse examples before models learn to extract robust user-specific patterns.

Persona information provides consistent benefits across few-shot regimes. Even at 32 shot, persona information yields substantial improvements (text: MAE 0.812 → 0.782, accuracy 0.421

→ 0.444; image: MAE 0.926 → 0.858, accuracy 0.392 → 0.428). This persistent benefit suggests that explicit persona information captures complementary signals to those learned from demonstration examples. Drawing parallels to recommender systems literature, our few-shot approach is analogous to item-based recommendation, while persona prompting resembles natural-language-based recommendation [See Sanner et al. (2023) for an overview]. Our results contribute to this line of research by demonstrating the potential value of hybrid approaches: while past behavior reveals specific preferences, persona information provides a broader context that may not be readily inferable from a limited set of behavioral examples.

Image few-shot prompting scales worse than text, despite zero-shot advantages. While image inputs yield the best zero-shot performance, they exhibit both steeper initial performance degradation and more limited few-shot scaling compared to text inputs. Despite showing consistent improvements with additional demonstrations, image performance does not surpass the zero-shot level even at 32 shots. This pattern likely reflects both the increased complexity of visual processing and limitations of current vision-language models in few-shot learning scenarios. Although images contain the complete information available to human annotators, current VLMs appear unable to fully leverage this rich visual information in few-shot contexts, suggesting an area for future work.

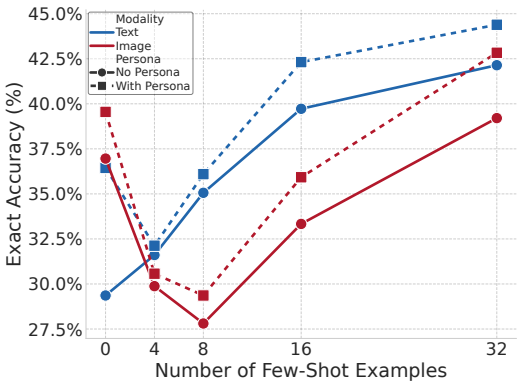


Figure 3: Few-shot learning performance, measured by exact match accuracy (%), as a function of the number of few-shot examples (0, 4, 8, 16, and 32).

Model	Input	Evaluation Metrics MAE↓	Acc↑	±1 Acc↑
A. Vision-Language Models				
Gemini-1.5 Pro	T	1.03	29.36	74.44
	I	0.94	36.96	77.03
	T+P	0.91	36.44	78.76
	I+P	0.84	39.55	82.04
GPT-4o	T	0.98	31.43	77.03
	I	0.91	35.41	79.97
	T+P	1.03	27.46	75.30
	I+P	0.89	36.79	79.45
Qwen2.5-VL-72B	T	0.99	32.82	78.76
	I	0.91	36.96	79.97
	T+P	0.92	34.72	80.31
	I+P	0.90	37.48	80.48
Llama-3.2-90B	T	1.14	33.51	71.16
	I	1.80	23.66	53.54
	T+P	0.90	36.44	81.86
	I+P	1.31	22.28	64.25
Grok-2	T	1.10	29.53	74.61
	I	1.04	34.02	74.09
	T+P	0.91	36.61	80.14
	I+P	0.90	37.13	81.52
B. Language-Only Models				
Llama-3.1-405B	T	0.98	34.20	78.07
	T+P	0.89	38.00	81.17
Llama-3.3-70B	T	1.16	31.61	71.16
	T+P	0.94	37.31	79.62

Input types: T=Text, I=Image, P=Persona
Metrics: MAE=Mean Absolute Error (↓ lower is better), Acc=Exact Accuracy, ±1 Acc=Within-One Accuracy (↑ higher is better)

Table 2: Zero-shot performance comparison across input modalities and models.

8 Conclusion

We introduce iNews, a novel dataset capturing individualized affective responses to news content, demonstrating that annotator characteristics could explain 15.2% of response variance - higher than existing datasets. In our case study, we find that incorporating persona information consistently improves prediction accuracy (up to 7% gains). We observe the early ascent phenomenon, where few-shot performance initially drops below zero-shot levels before recovering with additional examples. Notably, even at 32-shot (achieving 44.4% accuracy), persona information continues to provide benefits, suggesting that explicit modeling of individual characteristics offers complementary signals to behavioral demonstrations. We also identify a modality gap: while image inputs excel in zero-shot settings, they show limited few-shot scaling compared to text inputs. Through its rich annotations and persona variables, iNews advances research in personalization, subjectivity, and affective computing, while providing new opportunities for studying human behavioral simulation.

9 Limitations

Annotator and Sample Scope While iNews achieves substantial demographic and geographic diversity, with annotators from a majority of UK postcodes and various backgrounds, our sample most likely does not constitute a representative sample of the UK population or Facebook users. Nonetheless, iNews provides rich insights into how different demographic groups respond to news content. Our focus on UK-based annotators and UK news outlets necessarily constrains the applicability of our findings to other cultural, political, and media ecosystems. Future work should extend this work to diverse global contexts to build a more comprehensive understanding of news perception across different populations.

Methodological Considerations Our emotion measurement framework, while grounded in established psychological constructs (VAD, Plutchik), may not capture the full complexity of emotional responses to news content. Like most studies in this domain, we rely on self-reported emotions, which can be subject to various biases such as social desirability. Future work could strengthen the validity of these measurements by incorporating physiological measures (e.g., skin conductance, facial expressions) or triangulating multiple measurement approaches.

Platform Coverage. We focused our data collection on Facebook posts, as Facebook remains the dominant social media news source in the UK as of 2024 (Ofcom, 2024b). While platform-specific effects may exist, our findings provide valuable insights into how users engage with news on a major distribution channel. Future work could extend this analysis to other platforms to understand platform-specific effects.

Data Quality Although we implemented multiple quality control measures (attention checks, Captcha verification) and used Prolific's platform, which claims to provide 100% genuine human participants[4], we cannot completely rule out the possibility of AI-generated responses. Our modeling results support the high quality of the collected annotations, though as with any large-scale annotation effort, maintaining perfect attention throughout all responses cannot be guaranteed.

[4]https://www.prolific.com/participant-pool

10 Ethical Considerations

This study was conducted with the approval of our institutional ethics review board. All annotators provided informed consent before participation and were compensated fairly according to UK National Living Wage. No personally identifiable information was collected in our dataset. To protect participant privacy, we paraphrase open-ended responses quoted in this paper while preserving their essential meaning. During data collection, Prolific IDs were temporarily used to link annotations with persona data, but the publicly released dataset will contain only newly generated, anonymized participant IDs. Given our focus on UK-based annotators and news sources, we recognize the inherent limitations in global generalizability. However, we made conscious efforts to ensure demographic diversity within our annotator pool through Prolific's stratified sampling features. We acknowledge that emotional responses to news can be culturally specific and have thoroughly documented our annotator demographics to enable future researchers to account for potential demographic skews in their analyses.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *ICML 2024 Workshop on In-Context Learning*.

Allen Azizian, Todd D. Watson, Muhammad A. Parvaz, and Nancy K. Squires. 2006. Time course of processes underlying picture and word evaluation: An event-related potential approach. *Brain Topography*, 18(3):213–222.

M. T. Bastos. 2016. Digital journalism and tabloid journalism. In B. Franklin and S. Eldridge, editors, *The Routledge Companion to Digital Journalism Studies*. Routledge. This is an Accepted Manuscript of a book chapter published by Routledge in The Routledge Companion to Digital Journalism Studies on 20 Oct 2016, available online: https://www.routledge.com/The-Routledge-Companion-to-Digital-Journalism-Studies/Franklin-Eldridge-II/p/book/9781138887961.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Mattan S. Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56):2815.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 2007. Affective norms for english text (anet): Affective ratings of text and instruction manual. Technical report D-1, University of Florida, Gainesville, FL.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Susan T Charles and Laura L Carstensen. 2010. Social and emotional aging. *Annual review of psychology*, 61:383–409.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

John R Crawford and Julie D Henry. 2004. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265.

Natascha de Hoog and Peter Verboon. 2020. Is the news making us unhappy? the influence of daily news exposure on emotional states. *British Journal of Psychology*, 111(2):157–173.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

William Fleeson and Eranda Jayawickreme. 2015. Whole trait theory. *Journal of Research in Personality*, 56:82–92. Integrative Theories of Personality.

Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

International Press Telecommunications Council. 2024. Media topics: Subject taxonomy for the media - the successor to the subject codes. `https://iptc.org/standards/media-topics/`. Accessed: 2024-12-30.

Aarti Iyer and Julian Oldmeadow. 2006. Picture this: emotional and political responses to photographs of the kenneth bigley kidnapping. *European Journal of Social Psychology*, 36(5):635–647.

Ann M Kring and Albert H Gordon. 1998. Sex differences in emotion: expression, experience, and physiology. *Journal of personality and social psychology*, 74(3):686.

Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. 2017. Introducing the open affective standardized image set (oasis). *Behavior Research Methods*, 49(2):457–470.

Kenza Lamot. 2022. What the metrics say. the softening of news on the facebook pages of mainstream media outlets. *Digital Journalism*, 10(4):517–536.

Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 1997. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology*, 66(Volume 66, 2015):799–823.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2024. TopViewRS: Vision-language models as top-view spatial reasoners.

In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807, Miami, Florida, USA. Association for Computational Linguistics.

Ziqian Lin and Kangwook Lee. 2024. Dual operating modes of in-context learning. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

George E Marcus, W Russell Neuman, and Michael MacKuen. 2000. *Affective intelligence and political judgment*. University of Chicago Press.

Cameron Martel, Gordon Pennycook, and David G. Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):47.

Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology.* the MIT Press.

Batja Mesquita and Nico H Frijda. 1992. Cultural variations in emotions: a review. *Psychological bulletin*, 112(2):179.

Meta AI. 2024a. Llama 3.2: Revolutionizing edge AI and vision with open multimodal models. Accessed: 2025-01-15.

Meta AI. 2024b. Llama 3.3: Model Cards and Prompt Formats. Accessed: 2025-01-15.

Meta AI. 2024c. Meta Llama 3.1: Pushing the Boundaries of AI Research. Accessed: 2025-01-15.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Nic Newman, David Levy, and R Nielsen. 2015. *Reuters Institute digital news report 2015: Tracking the future of news*. Reuters Institute for the Study of Journalism.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Ofcom. 2024a. Adults news consumption survey - online questionnaire. Accessed: 2025-01-15.

Ofcom. 2024b. News consumption in the uk: 2024. Accessed: 2025-01-15.

Mary Beth Oliver. 2002. Individual differences in media effects. In *Media effects*, pages 517–534. Routledge.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Thomas E. Powell, Hajo G. Boomgaarden, Knut De Swert, and Claes H. de Vreese. 2015. A clearer picture: The contribution of visuals and text to framing effects. *Journal of Communication*, 65(6):997–1017.

David Preece, Rodrigo Becerra, and Guillermo Campitelli. 2018. Assessing emotional reactivity: Psychometric properties of the perth emotional reactivity scale and the development of a short form. *Journal of Personality Assessment*.

QwenLM. 2025. Qwen2.5-VL: A Vision-Language Model. Accessed: 2025-01-15.

Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.

Reuters Institute. 2024. Digital news report 2024 uk questionnaire. Accessed: 2025-01-15.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language- and item-based preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 890–896, New York, NY, USA. Association for Computing Machinery.

Roger N. Shepard. 1967. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1):156–163.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*.

Stuart Soroka, Patrick Fournier, and Lilach Nir. 2019. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116(38):18888–18892.

Hannah Sterz, Jonas Pfeiffer, and Ivan Vulić. 2024. Dare: Diverse visual question answering with robustness evaluation. *arXiv preprint arXiv:2409.18023*.

Justin Storbeck and Gerald L Clore. 2008. Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and personality psychology compass*, 2(5):1824–1843.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Patti M Valkenburg and Jochen Peter. 2013. The differential susceptibility to media effects model. *Journal of communication*, 63(2):221–243.

xAI. 2025. Grok-2. Accessed: 2025-01-15.

Jonathan R. Zadra and Gerald L. Clore. 2011. Emotion and perception: the role of affective information. *WIREs Cognitive Science*, 2(6):676–685.

A Appendix

A.1 Full Persona Variables

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

Age		
18-24 years old	23	7.9 %
25-34 years old	89	30.6 %
35-44 years old	77	26.5 %
45-54 years old	49	16.8 %
55-64 years old	37	12.7 %
65+ years old	16	5.5 %
Sex		
Male	152	52.2 %
Female	139	47.8 %
LGBTQ+ Self-Identification		
Yes	253	86.9 %
No	38	13.1 %
Ethnicity (Simplified)		
White	239	82.1 %
Mixed	20	6.9 %
Asian	17	5.8 %
Black	15	5.2 %
Personal Income (GBP)		
Less than £10,000	57	19.6 %
£10,000 - £19,999	60	20.6 %
£20,000 - £29,999	72	24.7 %
£30,000 - £39,999	49	16.8 %
£40,000 - £49,999	24	8.25 %
£50,000 - £59,999	12	4.12 %
£60,000 - £69,999	6	2.06 %
£70,000 - £79,999	3	1.03 %
£80,000 - £89,999	5	1.72 %
£90,000 - £99,999	2	0.687 %
More than £150,000	1	0.344 %
Highest Education Level Completed		
No formal qualifications	3	1.03 %
Secondary education (e.g. GED/GCSE)	24	8.25 %
High school diploma/A-levels	49	16.8 %
Technical/community college	38	13.1 %
Undergraduate degree (BA/BSc/other)	124	42.6 %
Graduate degree (MA/MSc/MPhil/other)	49	16.8 %
Doctorate degree (PhD/other)	4	1.37 %
Are you a student?		
Yes	34	11.7 %
No	246	84.5 %
DATA EXPIRED	11	3.8 %
Employment Status		
Due to start a new job within the next month	3	1.03 %
Other	6	2.06 %

Continued on next page

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

DATA EXPIRED	12	4.12 %
Unemployed (and job seeking)	13	4.47 %
Not in paid work (e.g. homemaker, retired or disabled)	36	12.4 %
Part-Time	50	17.2 %
Full-Time	171	58.8 %
Nationality (UK)		
England	232	79.7 %
Scotland	30	10.3 %
Wales	18	6.19 %
Northern Ireland	11	3.78 %
Political Leaning		
Centre	118	40.5 %
Right	84	28.9 %
Left	82	28.2 %
DATA EXPIRED	7	2.4 %
How interested, if at all, would you say you are in politics and the news?		
Not at all interested	3	1.03 %
Not very interested	30	10.3 %
Somewhat interested	113	38.8 %
Very interested	102	35.1 %
Extremely interested	43	14.8 %
In the past week, on average, how much time per day did you spend consuming news from all sources (online, TV, print, radio, etc.)?		
Less than 15 minutes	14	4.81 %
15 minutes to less than 30 minutes	45	15.5 %
30 minutes to less than 1 hour	94	32.3 %
1 hour to less than 2 hours	77	26.5 %
2 hours or more	61	21.0 %
How confident are you in your ability to distinguish between reliable and unreliable news sources?		
Not at all confident	3	1.03 %
Slightly confident	37	12.7 %
Moderately confident	109	37.5 %
Quite confident	123	42.3 %
Completely confident	19	6.5 %
How often do you fact-check news stories you come across?		
Never	14	4.81 %
Rarely	55	18.9 %
Often	84	28.9 %
Sometimes	132	45.4 %
Always	6	2.06 %
When reading a news article, how often do you consider the author's potential biases or agenda?		
Never	3	1.03 %
Rarely	27	9.28 %

Continued on next page

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

Often	112	38.5 %
Sometimes	105	36.1 %
Always	44	15.1 %
To what extent do you agree or disagree with the following statement: I often get emotionally affected by the news I read.		
Strongly disagree	6	2.06 %
Disagree	47	16.2 %
Neither agree nor disagree	62	21.3 %
Agree	157	54.0 %
Strongly agree	19	6.53 %
PLEASE SELECT ALL THAT APPLY: Which of the following platforms do you use for news nowadays?		
Radio	270	92.8 %
Website	263	90.4 %
Social media	217	74.6 %
Television	203	69.8 %
Word of mouth	164	56.4 %
Podcasts	88	30.2 %
Print newspaper / magazines	75	25.8 %
I don't	0	0 %
Which of the following social media sites do you use on a regular basis (at least once a month)? Choose any that apply.		
Youtube	243	83.5 %
Facebook	218	74.9 %
Instagram	171	58.8 %
Twitter	150	51.5 %
Reddit	119	40.9 %
Linkedin	97	33.3 %
TikTok	69	23.7 %
Pinterest	62	21.3 %
Snapchat	48	16.5 %
Google Plus	15	5.15 %
Tumblr	13	4.47 %
Meetup	9	3.09 %
Flickr	5	1.72 %
Medium	3	1.03 %
VK	3	1.03 %
Vine.co	2	0.687 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - BBC		
Don't know	0	0 %
Very untrustworthy	43	14.4 %
Untrustworthy	37	12.3 %
Neither trustworthy nor untrustworthy	28	9.3 %
Trustworthy	140	46.7 %
Very trustworthy	52	17.3 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Financial Times		

Continued on next page

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

Don't know	20	6.87 %
Very untrustworthy	11	3.78 %
Untrustworthy	17	5.84 %
Neither trustworthy nor untrustworthy	49	16.8 %
Trustworthy	156	53.6 %
Very trustworthy	38	13.1 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Guardian		
Don't know	13	4.47 %
Very untrustworthy	24	8.25 %
Untrustworthy	23	7.90 %
Neither trustworthy nor untrustworthy	55	18.9 %
Trustworthy	152	52.2 %
Very trustworthy	24	8.25 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Times & Sunday Times		
Don't know	19	6.53 %
Very untrustworthy	17	5.84 %
Untrustworthy	33	11.3 %
Neither trustworthy nor untrustworthy	75	25.8 %
Trustworthy	130	44.7 %
Very trustworthy	17	5.84 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Independent		
Don't know	15	5.15 %
Very untrustworthy	13	4.47 %
Untrustworthy	32	11.0 %
Neither trustworthy nor untrustworthy	79	27.1 %
Trustworthy	137	47.1 %
Very trustworthy	15	5.15 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - Sky		
Don't know	7	2.41 %
Very untrustworthy	36	12.4 %
Untrustworthy	59	20.3 %
Neither trustworthy nor untrustworthy	76	26.1 %
Trustworthy	98	33.7 %
Very trustworthy	15	5.15 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Economist		
Don't know	29	9.97 %
Very untrustworthy	10	3.44 %
Untrustworthy	24	8.25 %
Neither trustworthy nor untrustworthy	69	23.7 %
Trustworthy	135	46.4 %
Very trustworthy	24	8.25 %

Continued on next page

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Telegraph		
Don't know	13	4.47 %
Very untrustworthy	24	8.25 %
Untrustworthy	47	16.2 %
Neither trustworthy nor untrustworthy	91	31.3 %
Trustworthy	100	34.4 %
Very trustworthy	16	5.50 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Metro		
Don't know	18	6.19 %
Very untrustworthy	39	13.4 %
Untrustworthy	64	22.0 %
Neither trustworthy nor untrustworthy	111	38.1 %
Trustworthy	54	18.6 %
Very trustworthy	5	1.72 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - GB News		
Don't know	17	5.84 %
Very untrustworthy	91	31.3 %
Untrustworthy	72	24.7 %
Neither trustworthy nor untrustworthy	69	23.7 %
Trustworthy	36	12.4 %
Very trustworthy	6	2.06 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Daily Mail		
Don't know	6	2.06 %
Very untrustworthy	125	43.0 %
Untrustworthy	81	27.8 %
Neither trustworthy nor untrustworthy	47	16.2 %
Trustworthy	27	9.28 %
Very trustworthy	5	1.72 %
How trustworthy or untrustworthy do you rate the news reported by the following media organizations? - The Mirror		
Don't know	8	2.75 %
Very untrustworthy	101	34.7 %
Untrustworthy	89	30.6 %
Neither trustworthy nor untrustworthy	57	19.6 %
Trustworthy	33	11.3 %
Very trustworthy	3	1.03 %
Cognitive Reflection Test - Number of Correct Answers		
0	58	19.9 %
1	44	15.1 %
2	67	23.0 %
3	122	41.9 %

Continued on next page

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

Extraversion		
1	37	12.7 %
1.5	24	8.25 %
2	43	14.8 %
2.5	49	16.8 %
3	59	20.3 %
3.5	17	5.84 %
4	34	11.7 %
4.5	16	5.50 %
5	12	4.12 %
Agreeableness		
1	6	2.06 %
1.5	8	2.75 %
2	23	7.90 %
2.5	24	8.25 %
3	59	20.3 %
3.5	59	20.3 %
4	42	14.4 %
4.5	44	15.1 %
5	26	8.93 %
Conscientiousness		
1.5	4	1.37 %
2	15	5.15 %
2.5	12	4.12 %
3	41	14.1 %
3.5	54	18.6 %
4	59	20.3 %
4.5	44	15.1 %
5	62	21.3 %
Neuroticism		
1	30	10.3 %
1.5	23	7.90 %
2	34	11.7 %
2.5	36	12.4 %
3	52	17.9 %
3.5	27	9.28 %
4	43	14.8 %
4.5	28	9.62 %
5	18	6.19 %
Openness		
1	1	0.344 %
1.5	5	1.72 %
2	18	6.19 %
2.5	24	8.25 %
3	51	17.5 %
3.5	61	21.0 %
4	47	16.2 %
4.5	43	14.8 %
5	41	14.1 %

Continued on next page

Table 3: **Full Persona Variables Breakdowns.** Counts and percentages of participants by persona variables. Big-Five Personality traits are assessed using the BFI-10 scale (Rammstedt and John, 2007). For the Perth Emotional Reactivity Scale (Preece et al., 2018), we include one question each from four dimensions (negative-activation, negative-intensity, positive-activation, positive-intensity) due to questionnaire length constraint (Preece et al., 2018). Additionally, we include net positive scores calculated from ten items on the Positive and Negative Affect Schedule (Crawford and Henry, 2004). Our dataset includes a wide range of persona variables at both group and individual levels.

Perth Emotional Reactivity Scale - Positive Activation: Please score the following statements according to how much they apply or do not apply to you. - I tend to get happy very easily.		
Very unlike me	24	8.25 %
Somewhat unlike me	68	23.4 %
Neither like or unlike me	67	23.0 %
Somewhat like me	104	35.7 %
Very like me	28	9.62 %
Perth Emotional Reactivity Scale - Positive Intensity: Please score the following statements according to how much they apply or do not apply to you. - I experience positive mood very strongly.		
Very unlike me	14	4.81 %
Somewhat unlike me	48	16.5 %
Neither like or unlike me	63	21.6 %
Somewhat like me	130	44.7 %
Very like me	36	12.4 %
Perth Emotional Reactivity Scale - Negative Activation: Please score the following statements according to how much they apply or do not apply to you. - I tend to get disappointed very easily.		
Somewhat like me	101	34.7 %
Somewhat unlike me	78	26.8 %
Neither like or unlike me	47	16.2 %
Very like me	40	13.7 %
Very unlike me	25	8.59 %
Perth Emotional Reactivity Scale - Negative Intensity: Please score the following statements according to how much they apply or do not apply to you. - My negative feelings feel very intense.		
Somewhat like me	94	32.3 %
Somewhat unlike me	60	20.6 %
Very like me	55	18.9 %
Neither like or unlike me	49	16.8 %
Very unlike me	33	11.3 %
Positive and Negative Affect Schedule - Net Positive Score	**Count**	**Percentage**
< -10	1	0.344 %
-10 to -5	1	0.344 %
-5 to 0	23	7.90 %
0 to 5	77	26.5 %
5 to 10	112	38.5 %
> 10	77	26.5 %

A.2 Pilot Study: Stimulus Modality Selection

We conducted a pilot study to determine whether to use full Facebook news post screenshots or text-only headlines as stimuli. This decision involves several trade-offs. Text-only stimuli are simpler to process with current language models and sufficient for many breaking news posts that use generic images. However, full screenshots offer greater ecological validity, as social media users typically encounter both text and images simultaneously. Prior research suggests that visual stimuli are processed more rapidly than text (Azizian et al., 2006) and are more memorable (Shepard, 1967), though current open-source vision-language models still face significant performance and robustness challenges (Li et al., 2024; Sterz et al., 2024).

To empirically inform this decision, we collect annotations from 40 UK-based participants (20 per condition) for 10 paired news posts from March 2024, present either as text-only headlines or full screenshots. Participants rated valence, arousal, and dominance (VAD) and provide discrete emotion categories.

We then analyze the aggregated ratings across all 10 posts for each condition. Table 4 presents the descriptive and inferential statistics for the dimensional emotions (VAD). Mann-Whitney U tests indicate significant differences in valence ($p = 0.016$) and dominance ($p = 0.019$), though with small effect sizes (RBC ranging from -0.032 to -0.136). For discrete emotions, a chi-square test indicates marginally significant differences in emotion distribution between conditions ($\chi^2 = 14.93, p = 0.060$). We additionally visually show the distribution of VAD and discrete ratings in Figures 4 and 5.

The distribution patterns of VAD and discrete ratings are visualized in Figures 4 and 5. While VAD distributions remain broadly similar across conditions, the image condition elicits more negative valence ratings and more neutral dominance ratings. The differences in discrete emotion ratings are more noticeable, with substantially fewer neutral emotions reported in the image condition. We interpret this as evidence that images help disambiguate emotional content - since the image condition includes both visual and textual information, it may provide richer context for emotional interpretation.

Based on these findings and theoretical considerations, we decide to use full screenshots for our main study. This choice is driven by observed differences in emotional annotations, the ecological validity of multimodal news consumption on social media, and the additional contextual information provided by images. While current vision-language models face technical limitations, we anticipate rapid advancement in multimodal processing capabilities and prioritize capturing more naturalistic news consumption experiences over immediate computational convenience.

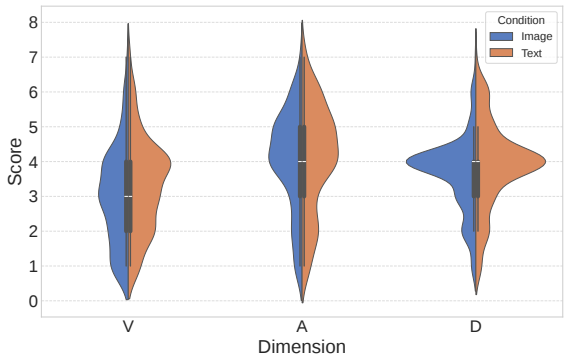


Figure 4: Distribution of VAD scores across modality conditions

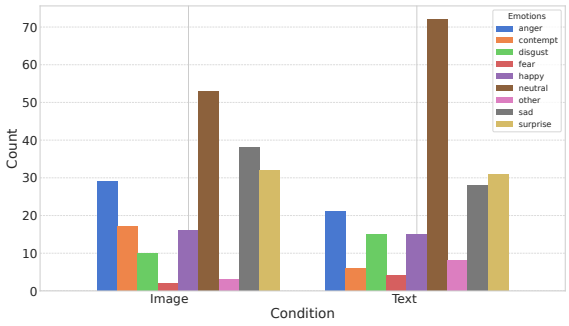


Figure 5: Distribution of discrete emotions across modality conditions

A.3 Geographic Representation of Annotators

We present the geographic distribution of annotators across UK postcode areas in Figure 6. Our dataset includes annotators from 97 of the 124 postcode areas in the UK, demonstrating broad geographical coverage. To assess the representativeness of our sample, we compute the Pearson correlation coefficient between the number of annotators per postcode area and the corresponding 2011 Census population figures. The resulting correlation of 0.662 indicates a moderate positive relationship between population density and annotator distribution. To further quantify geographic representation, we calculate a representativeness ratio for each postcode area by dividing the percentage of annotators in each area by the percentage of the UK population in that same area. The mean ratio of 1.26 indicates that most areas are well-represented, often exceeding proportional representation. While there is some variation (standard deviation of 1.05 and median of 0.95), the overall distribution suggests we achieved strong geographic diversity in our sample.

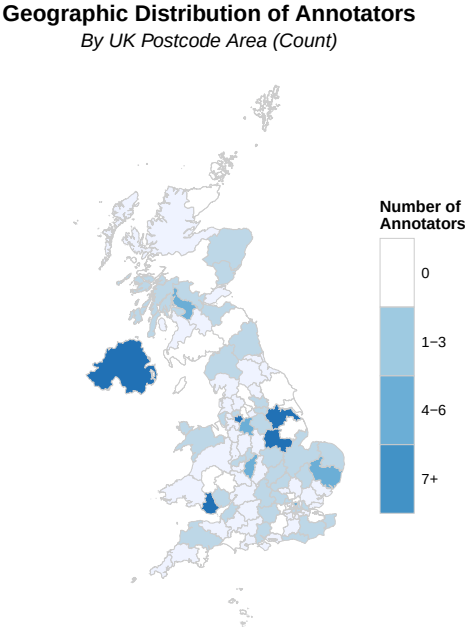


Figure 6: Geographic distribution of annotators across UK postcode areas.

	Image		Text-only		Statistics	
Dimension	M (SD)	Mdn	M (SD)	Mdn	U	p-value
Valence	3.03 (1.40)	3.0	3.35 (1.36)	3.0	17,272.5	.016*
Arousal	4.02 (1.42)	4.0	4.09 (1.53)	4.0	19,366.0	.574
Dominance	3.66 (1.04)	4.0	3.90 (1.19)	4.0	17,529.5	.019*

Table 4: Comparison of affective response between image and text-only conditions.* denotes p < .05.

Page Name	Mean V	SD V	Mean A	SD A	Mean D	SD D	Count	Discrete %
BBC News	3.45	1.08	4.42	1.14	3.72	1.01	498	sad: 24.6%, neutral: 22.3%
The Independent	3.49	0.99	4.28	1.12	3.82	0.85	339	neutral: 34.6%, sad: 17.5%
Daily Mail	3.09	0.97	4.40	1.21	3.61	1.00	305	sad: 28.4%, neutral: 26.7%
The Mirror	3.59	0.97	4.03	1.19	3.85	0.83	240	neutral: 40.1%, sad: 18.7%
Metro	3.44	0.98	4.19	1.19	3.77	0.91	213	neutral: 35.6%, sad: 19%
The Sun	3.56	1.03	4.07	1.27	3.88	0.87	213	neutral: 35.9%, sad: 17%
The Telegraph	3.37	1.15	4.39	1.12	3.79	1.04	190	neutral: 29%, sad: 16.2%
Daily Express	3.70	0.91	3.85	1.17	3.89	0.70	141	neutral: 49.5%, sad: 13.3%
The Guardian	3.68	1.06	4.21	1.19	3.88	0.96	136	neutral: 30.5%, sad: 20.7%
The Economist	3.61	0.99	4.35	1.08	3.74	0.93	126	neutral: 39.3%, happy: 11%
Daily Star	3.67	1.10	4.12	1.16	4.01	0.79	109	neutral: 39.5%, happy: 13.3%
The i Paper	3.32	1.17	4.42	1.10	3.72	1.09	89	neutral: 26.3%, sad: 19.1%
ITV News	3.35	1.13	4.45	1.10	3.70	1.10	69	sad: 22.5%, neutral: 21.9%
The Times and The Sunday Times	3.30	1.23	4.55	1.17	3.71	1.15	50	neutral: 23.8%, anger: 17.2%
LADbible	3.67	0.98	4.46	1.18	3.81	1.03	49	surprise: 24.7%, neutral: 21.2%
Sky News	2.90	1.04	4.74	1.15	3.50	1.10	41	sad: 26.6%, neutral: 19.6%
GB News	3.49	1.10	4.09	1.12	3.90	0.95	40	neutral: 31.9%, sad: 15.2%
Reuters UK	3.71	1.23	4.27	1.21	3.84	1.11	24	neutral: 29.5%, surprise: 18%
LBC	3.40	1.12	4.30	1.16	3.78	1.02	19	sad: 27.6%, neutral: 19.4%
Financial Times	3.95	1.18	4.28	1.20	3.87	1.14	8	neutral: 39.5%, surprise: 16.3%

Table 5: Distribution of valence, arousal, and dominance and discrete emotion labels by outlet.

A.4 Extended Descriptive Analysis

A.4.1 Additional Annotation Distributions Analysis

We show the distributions of the collected annotation variables in Figure 7. In addition to the discussions in Section 4, regarding relevance (Figure 7e), almost half of the annotations (44%) indicate "Not at all" relevant, with only 3.8% marked as "extremely relevant." For sharing inclination (Figure 7f), the distribution is even more skewed, with 54.5% of the annotations indicating "very unlikely" to share.

The majority of annotations (52.3%, Figure 7g) suggest that both the text and image significantly influence emotional reactions to news headlines. In contrast, approximately a third (36.7%) highlight the text alone as the primary factor. This indicates the importance of considering both the image and the text when modeling affective responses to news headlines on social media, rather than focusing solely on one or the other.

A.4.2 Outlet-level Analysis

To examine the traditional distinction between broadsheet and tabloid publications[5], we conduct Welch's t-tests comparing their affect scores. Interestingly, we find no significant differences in valence ($p = 0.83$) or dominance ($p = 0.64$) between the two types of outlets. However, there is a marginally significant difference in arousal ($p = 0.052$), with broadsheet publications eliciting slightly higher arousal responses ($M = 4.34$) compared to tabloids ($M = 4.09$). While the digital transformation of news media might have blurred many traditional distinctions between tabloids and broadsheets, these findings suggest that different editorial standards may still influence readers' affective responses, particularly in terms of emotional arousal.

A.4.3 Relationship Between Arousal and Valence

We calculate the average valence and arousal for each headline and present the results in Figure 9. Point opacity indicates overlapping points, suggesting a higher density of images. The distribution follows a V-shaped pattern, where arousal levels are high at both low and high extremes of valence, and this pattern aligns with established findings in affective science (Lang et al., 1997; Kurdi et al., 2017). However, our data also present notable deviations. Specifically, we observe a higher concentration of headlines exhibiting elevated arousal levels (above 6) in both the first and second quadrants (low valence/high arousal and high valence/high arousal, respectively). This concentration is particularly pronounced in the second quadrant, characterized by very low valence and very high arousal. We also see a concentration of density along the central region, around arousal ≈ 4 and valence ≈ 4, with a slight skew towards the upper-left quadrant. Finally, the overall distribution in our dataset encompasses a broader region of the valence-arousal space compared to that of Kurdi et al. (2017). We hypothesize that this discrepancy arises from the inherently negative nature of news headlines, in contrast to the more emotionally diverse stimuli typically employed in prior studies comprising images of scenes and objects.

A.5 Inter-annotator agreement

We measure Krippendorff's α for each of the annotated variables and present the results in Table 6. Among the core emotional dimensions, valence shows moderate agreement ($\alpha = 0.468$), while arousal and dominance exhibit lower agreement levels ($\alpha = 0.145$ and $\alpha = 0.203$, respectively). Discrete emotion categories demonstrate comparable levels of agreement. The auxiliary variables—modality importance, relevance, and sharing intent—show particularly low agreement ($\alpha < 0.1$).

Dimension	Krippendorff's α
Valence	0.468
Arousal	0.145
Dominance	0.203
Discrete Emotions	0.202
Modality Importance	0.083
Relevance	0.079
Sharing Intent	0.057

Table 6: Inter-annotator agreement measured by Krippendorff's α for continuous dimensions (V/A/D), discrete emotions, and auxiliary variables.

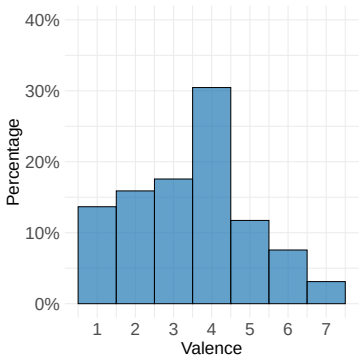
These findings align with previous research in emotion and affect annotation. The relatively low inter-annotator agreement is consistent with similar datasets (Strapparava and Mihalcea, 2007; Busso et al., 2008; Demszky et al., 2020; Oberländer et al., 2020), and the pattern of higher agreement for valence compared to arousal and dominance mirrors observations in prior work (Busso et al., 2008; Buechel and Hahn, 2017). These low agreement levels highlight a crucial insight: emotional responses to news content are inherently subjective and individualized. This observation strengthens our argument for modeling personalized affective responses rather than pursuing consensus annotations.

A.6 Topic Classification Details

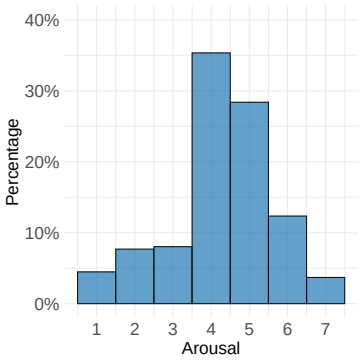
We apply the following prompt in JSON mode with the gemini-1.5-pro endpoint.

```
You are an expert news content analyst.

**Task 1:**

Your task is to describe the IMAGE in
↪  this Facebook news post and how it
↪  works together with the post's text.
↪  Focus on:

- What's shown in the image
- How the image and text complement or
↪  contrast with each other
- Any visual elements that particularly
↪  grab attention (e.g., graphics,
↪  colors, composition)

**Task 2:**

Focus only on the attached the IMAGE of
↪  the news post to select the SINGLE
↪  most appropriate category:

Categories and definitions

**Task 3:**
```

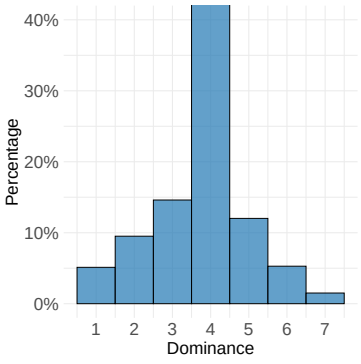
[5] We classify The Times, The Telegraph, The Guardian, The Independent, i Paper, and Financial Times as broadsheets, and The Sun, Daily Mail, Daily Express, The Mirror, and Daily Star as tabloids. All other outlets are categorized as "other". See Bastos (2016) for a comparison of broadsheet and tabloid newspapers, both historically and in the present day.



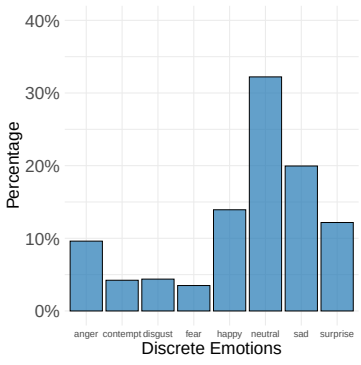
(a) Valence Distribution



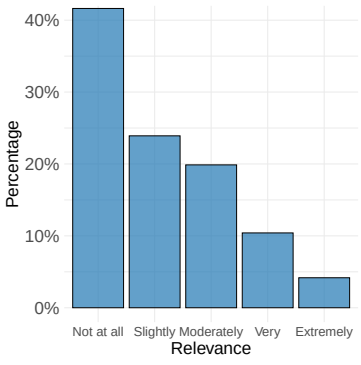
(b) Arousal Distribution



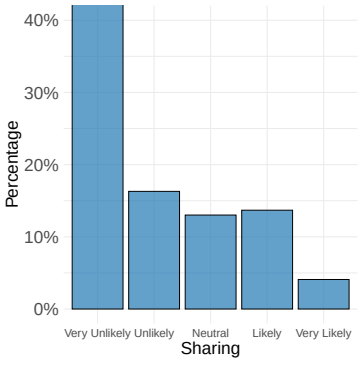
(c) Dominance Distribution



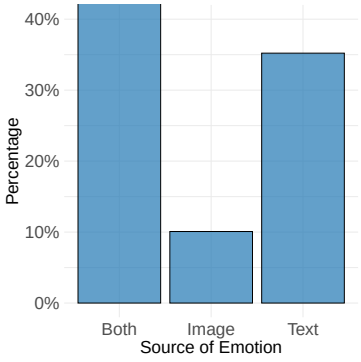
(d) Discrete Distribution



(e) Relevance Distribution



(f) Sharing Distribution



(g) Source Distribution

Figure 7: Distribution of Annotations

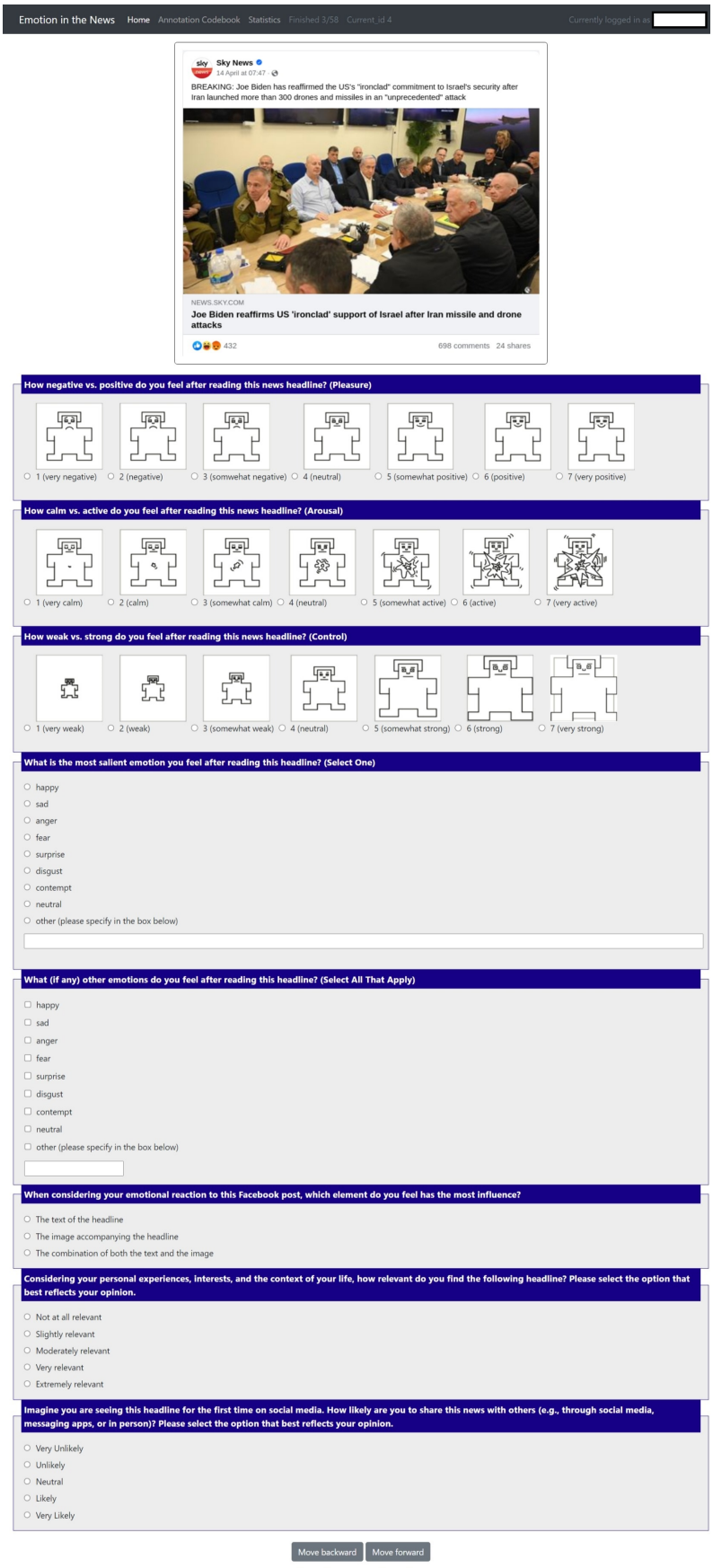


Figure 8: A screenshot of the annotation interface.

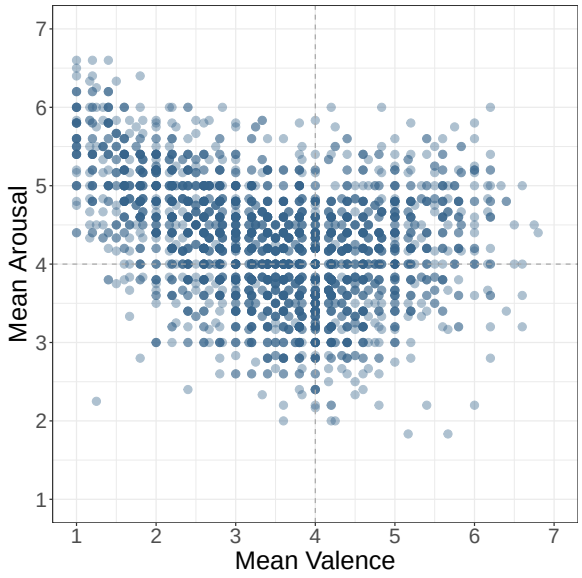


Figure 9: Distribution of affective responses to news posts in the valence-arousal space. Each point represents the mean arousal and valence ratings for a single headline, with darker regions indicating higher density of overlapping points.

```
Provide a confidence score (1-100)
↪  indicating how certain you are of
↪  your choice in Task 2.

**Task 4:**

Analyze both the text content and image
↪  to select the SINGLE most
↪  appropriate category and provide a
↪  confidence score (1-100):

Categories and definitions

- arts, culture, entertainment and
↪  media: All forms of arts,
↪  entertainment, cultural heritage
↪  and media
- conflict, war and peace: Acts of
↪  socially or politically motivated
↪  protest or violence, military
↪  activities, geopolitical conflicts,
↪  as well as resolution efforts
- crime, law and justice: The
↪  establishment and/or statement of
↪  the rules of behaviour in society,
↪  the enforcement of these rules,
↪  breaches of the rules, the
↪  punishment of offenders and the
↪  organisations and bodies involved in
↪  these activities
- disaster, accident and emergency
↪  incident: Man made or natural event
↪  resulting in loss of life or injury
↪  to living creatures and/or damage to
↪  inanimate objects or property
- economy, business and finance: All
↪  matters concerning the planning,
↪  production and exchange of wealth.
- education: All aspects of furthering
↪  knowledge, formally or informally
- environment: All aspects of
↪  protection, damage, and condition of
↪  the ecosystem of the planet earth
↪  and its surroundings.
- health: All aspects of physical and
↪  mental well-being
- human interest: Item that discusses
↪  individuals, groups, animals, plants
↪  or other objects in an emotional way
- labour: Social aspects, organisations,
↪  rules and conditions affecting the
↪  employment of human effort for the
↪  generation of wealth or provision of
↪  services and the economic support of
↪  the unemployed.
- lifestyle and leisure: Activities
↪  undertaken for pleasure, relaxation
↪  or recreation outside paid
↪  employment, including eating and
↪  travel.
- politics: Local, regional, national
↪  and international exercise of power,
↪  or struggle for power, and the
↪  relationships between governing
↪  bodies and states.
- religion: Belief systems, institutions
↪  and people who provide moral
↪  guidance to followers
- science and technology: All aspects
↪  pertaining to human understanding
↪  of, as well as methodical study and
↪  research of natural, formal and
↪  social sciences, such as astronomy,
↪  linguistics or economics
- society: The concerns, issues, affairs
↪  and institutions relevant to human
↪  social interactions, problems and
↪  welfare, such as poverty, human
↪  rights and family planning
- sport: Competitive activity or skill
↪  that involves physical and/or mental
↪  effort and organisations and bodies
↪  involved in these activities
- weather: The study, prediction and
↪  reporting of meteorological
↪  phenomena

**Task 5:**

Provide a confidence score (1-100)
↪  indicating how certain you are of
↪  your choice in Task 4.
```

A.7 Regression Analysis Details

This section provides additional details on the regression models used in the main text (Section 5), including full model specifications, results for additional models, and a discussion of variable importance.

Model Specifications and Estimation We employ linear mixed-effects models (LMMs) to analyze the influence of persona variables and other factors on annotators' arousal ratings. LMMs are appropriate for this analysis because they account for the nested structure of the data (multiple annotations per news post and per annotator) and allow for both

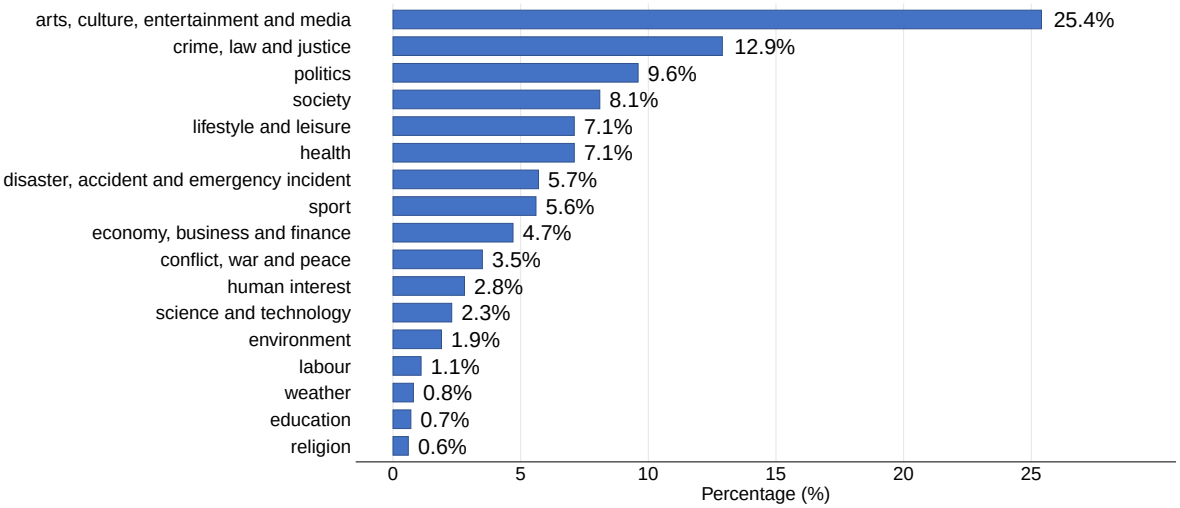


Figure 10: Distribution of news articles across topic categories.

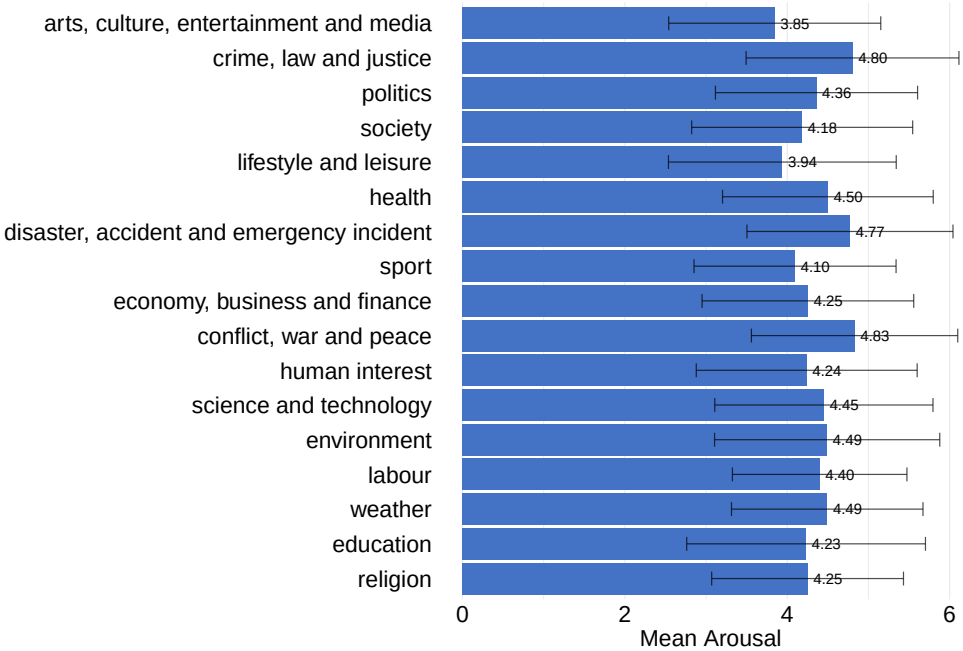


Figure 11: Average arousal scores by topic category, with error bars indicating one standard deviation from the mean.

fixed effects (e.g., persona variables) and random effects (e.g., individual differences between annotators and news posts). All models are estimated using the `lme4` package (Bates et al., 2015) in R. The dependent variable in all models is the annotator's arousal rating for a given news post (ranging from 1 to 7).

The following models are estimated in the main text in Section 5:

1. **Null Model:** Baseline model with only a random intercept for news text.

```
Arousal ~ 1 + (1 | Text)
```

2. **Persona Model:** Includes 47 persona variables as fixed effects and a random intercept for news text.

```
Arousal ~ PersonaVariables + (1 | Text)
```

   where `PersonaVariables` represents the full set of 47 persona variables.

3. **User Model:** Includes random intercepts for both news text and annotator ID.

```
Arousal ~ 1 + (1 | Text) + (1 | UserID)
```

Additional Models To explore the contributions of other contexual factors, we estimate these additional models:

4. **Outlet Model:** Adds news outlet as a fixed effect.

```
Arousal ~ PersonaVariables + Outlet
↪  + (1 | Text)
```

5. **Calibration Model:** Adds responses to the three calibration items as fixed effects.

```
Arousal ~ PersonaVariables + Calibration
↪  + (1 | Text)
```

6. **Topic Model:** Adds news post topic category as a fixed effect.

```
Arousal ~ PersonaVariables + Topic + (1
↪  | Text)
```

7. **All Model:** Combines all fixed effects from the Outlet, Calibration, and Topic models.

```
Arousal ~ PersonaVariables + Outlet +
↪  Calibration + Topic + (1 | Text)
```

8. **All + User Model:** Adds a random intercept for user ID to the All Model.

```
Arousal ~ PersonaVariables + Outlet +
↪  Calibration + Topic + (1 | Text) +
↪  (1|UserID)
```

Full Regression Results Table 7 presents the full results for all models, including marginal and conditional R^2 values, calculated using the method described by Nakagawa and Schielzeth (2013).

Variable Importance What are the most important persona variables? Is it more the case that some specific persona variables explain the vast majority of variance or is it rather spread out across all variables? To answer this question, we analyze the Persona model and calculated the Eta-squared (η^2), a commonly used measure representing the proportion of the total variance in the dependent variable accounted for by a given independent variable. The calculations are performed using the `effectsize` package (Ben-Shachar et al., 2020) in R.

Based on effect sizes, the individual contributions of the persona variables to explaining variance in arousal are generally modest. The majority of persona variables have small effect sizes, below 0.005. We show the top 10 persona variable with highest effect sizes in Table 8. Despite this overall trend, a subset of variables exhibit somewhat larger effect sizes. These included factors related to socioeconomic status, such as personal income ($\eta^2 = 0.010$) and education level ($\eta^2 = 0.008$), as well as employment status ($\eta^2 = 0.008$). Personality traits also demonstrate notable influence, particularly Agreeableness ($\eta^2 = 0.009$) and Neuroticism ($\eta^2 = 0.008$). Among media consumption patterns, television viewing habits stand out ($\eta^2 = 0.008$), while current emotional state also show meaningful effects ($\eta^2 = 0.007$).

These findings suggest that while the regression model as a whole demonstrates a reasonable ability to predict arousal (as indicated by the R^2 values discussed previously), the influence of individual persona variables is, for the most part, limited. The observed model fit likely stems from the cumulative effect of numerous variables with small individual contributions. This pattern aligns with the complex, multifaceted nature of affective responses to news content, where multiple personal characteristics interact to shape individual reactions (also see the quantitative interview analysis in Section A.8. The distributed nature of these effects underscores the importance of considering a broad spectrum of persona variables in modeling affective responses, rather than focusing on a limited set of characteristics.

Analysis of Content and Behavioral Effects While our analyses in the main paper focus on modeling individual differences through user-level variables, our dataset contains rich metadata about the content itself: news topics, headline image categories (see Section A.6), and source outlets. We also collected calibration data by having annotators respond to three standardized items from the ANET dataset. To understand the relative importance of these factors, we first evaluate their contributions separately (Outlet, Calibration, and Topic models) before combining them (All model).

The Outlet and Topic models, which incorporate static content features, achieve similar total explanatory power (conditional R^2) to the Persona model but with higher fixed-effect contributions (marginal R^2). This suggests these content-based features capture some of the variance previously attributed to random effects, without improving overall prediction. In contrast, the Calibration model shows higher total explanatory power ($R^2 = 0.328$ vs. 0.286), indicating that annotators' annotation behavior on the calibration items may potentially capture variance unexplained by our carefully selected persona variables.

The All model, despite incorporating numerous fixed effects, maintains approximately the same conditional R^2 as the Calibration model. However, it demonstrates a substantial shift in R^2 distribution, with marginal R^2 reaching 0.261—exceptional for annotator modeling in NLP (Hu and Collier, 2024). Notably, we achieve better conditional R^2 compared to the User model (random-effects only), which only includes random intercepts for text stimuli. This improvement likely stems from two factors: first, the inherently conservative nature of random-effects fitting, which employs regularization

Model	Fixed Effects	Random Effects	Variance Explained	
			Marginal R^2	Conditional R^2
Baseline Models				
Null	None	Text	0.000	0.131
User	None	Text + User	0.000	0.317
Individual Differences				
Persona	Persona	Text	0.152	0.286
Persona + User	Persona	Text + User	0.139	0.377
Content & Behavior				
Outlet	Persona + Outlet	Text	0.166	0.287
Calibration	Persona + Calibration	Text	0.193	0.328
Topic	Persona + Topics	Text	0.217	0.283
Combined Models				
All	Persona + All Above	Text	**0.261**	0.326
All + User	Persona + All Above	Text + User	0.239	**0.383**

Table 7: Regression analysis of affective arousal to news headlines. Models progress from baseline through increasingly complex specifications, incorporating individual differences (persona variables), content features (outlet, topic), and behavioral measures (calibration). Marginal R^2 shows variance explained by fixed effects alone, while conditional R^2 includes both fixed and random effects.

Parameter	Partial η^2
Personal Income (GBP)	0.010
Agreeableness	0.009
Neuroticism	0.008
Employment Status	0.008
Television	0.008
Current Emotional State (PANAS)	0.008
Highest Education Level Completed	0.008
News Trust: Independent	0.007
News Trust: Mirror	0.007
Extraversion	0.007

Table 8: Top 10 variables with the largest effect sizes (partial η^2) in the Persona Model

to prevent overfitting; and second, random effects' limitation in capturing structural information within the data. While random effects excel at modeling individual-level variation, they treat such variation as purely stochastic, potentially overlooking systematic patterns that our comprehensive set of fixed effects can capture. Our results demonstrate that affective responses to news content, though complex, exhibit structural patterns that can be systematically modeled through carefully selected persona variables, including demographic characteristics, psychological traits, and news consumption behaviors.

Analysis of User Random Effects Given the previous results, we then investigate whether adding user-level random effects benefits models with rich fixed effects. In theory, perfect fixed effects would eliminate the need for user-level random effects. In practice, however, adding user-level random effects improves model fit for both the Persona (Personal + User model) and All (All + User model) models, though with diminishing returns. The improvement is smaller for the All model ($\Delta = 0.057$) compared to the Persona model ($\Delta = 0.091$), suggesting we may be approaching a ceiling for random effects gains. This asymptotic behavior indicates that while better fixed effects reduce the potential contribution of random effects, our current setup has not yet exhausted all relevant fixed effect variables, leaving room for future data collection and modeling improvements.

Discussion of Regression Results Our findings connect to a fundamental question in psychology: do people's reactions come from who they are (their personality, beliefs, demographics) or from what they're responding to (in our case, the news content)? Our results indicate both, supporting an interactionist perspective (Mischel and Shoda, 1995) - person-level variables and stimulus (the news posts) both contribute meaningfully to explaining affective responses, with their combination yielding higher explanatory power.

The persistent benefit of including user-level random effects, even in our most comprehensive model ($\Delta R^2 = 0.057$), aligns with contemporary personality theory (Fleeson and Jayawickreme, 2015) which conceptualizes individual differences through density distributions. This framework suggests that while considerable behavioral variability exists within each individual, the parameters of these distributions may be stable across. In our case, this means that while a person's affective responses to news may vary substantially across different stories, their pattern of variation itself could be characteristic and predictable. This theoretical perspective helps explain why both fixed effects (capturing systematic individual differences) and random effects (accounting for person-specific response patterns) contribute uniquely to our model's predictive power.

A.8 Post-Annotation Questionnaire

To better understand how annotators approached the task and complement the quantitative analysis of persona variables, we

conduct a post-annotation qualitative study using a detailed questionnaire. The questionnaire is shown below. Following the questionnaire, we present an in-depth analysis of the responses for each question.

> **Q1:** How do you think your personal background (e.g., age, education, political views) influenced your emotional responses to the news headlines?
>
> **Q2:** Did you notice any patterns in the types of headlines that elicited stronger emotional responses from you? If so, what were they?
>
> **Q3:** How do you think your emotional responses to these headlines might differ from those of the general public?
>
> **Q4:** How do you think your media consumption habits (e.g., frequency, preferred sources) might have affected your responses to these Facebook news posts? Did you notice any differences in your responses to news posts from different sources or publishers?
>
> **Q5:** Reflecting on your experience annotating these news posts, what do you believe were the top 3-5 factors that most influenced your emotional responses? These could be related to the news content itself, your personal background, or external circumstances. For each factor, please briefly explain how you think it affected your reactions.
>
> **Q6:** Reflecting on your experience with this annotation task and how you typically consume news, are there any insights, observations, or personal reflections you'd like to share about how you engaged with and responded to the news posts, or anything else you'd like to share?

A.8.1 Q1

Regarding the influence of personal background (Q1), annotators demonstrate a keen awareness of how factors including age and lived experiences, political affiliations, educational background, media literacy and consumption habits and personal values shape their emotional processing of news. For instance, one annotator reflects on how their generation's experience during the cold war impacts their reactions to current events, stating that they "get this pit in my stomach when I read these stories" due to specific events experienced during their lifetime, which differs from the experiences of younger people. Another annotator emphasizes the impact of political views on their emotional responses, noting that they feel "really frustrated and annoyed" towards content that conflicted with their political ideology. These examples illustrate how personal history and deeply held beliefs create unique perspectives and biases, coloring readers' emotional engagement with the news. Additionally, many annotators report becoming desensitized due to constant exposure to negative news and recognized modern phenomena like clickbait. There is a notable awareness of how different news sources operate, with some annotators expressing inherent distrust of certain outlets.

A.8.2 Q2

When analyzing the types of headlines that elicit stronger responses (RQ2), we observe a clear distinction between content-driven and presentation-driven factors. Regarding content, annotators consistently identify news related to harm, suffering, and threats to vulnerable populations as powerful emotional triggers. One annotator's comment captures this pattern: "I really feel it more when the story is about people getting hurt, especially when it's kids or families." Contemporary societal issues also generate intense responses, with annotators citing topics such as COVID-19, immigration, healthcare systems, and international conflicts. Personal relevance emerges as another crucial content factor, with annotators responding more intensely to news that mirrors their experiences or aligns with their values. As one annotator puts it: "when it's something I've been through myself, or it reminds me of my own family, it really gets to me."

In terms of presentation, visual elements significantly influence emotional intensity. Multiple annotators report that headlines accompanied by images, especially those depicting suffering or tragedy, elicit stronger emotional reactions, with some finding certain visual content overwhelming. This finding validates our research design's inclusion of complete news post screenshots rather than headlines alone. Source credibility also shapes emotional engagement, with annotators expressing greater trust in established news sources (e.g., BBC) compared to social media, and demonstrating skepticism toward tabloids and sensationalized content.

A.8.3 Q3

In exploring potential differences between their responses and those of the "general public" (Q3), responses mention both universality and divergence. While there is acknowledgment of shared emotional ground, particularly regarding responses to tragedy, suffering, and social norm violations, these mentions are often qualified by extensive discussion of individual variations. A strong theme emerges around the recognition of response variability, with one annotator articulating that "everyone's got their own way of feeling about things - you can't expect two people to react exactly the same." Participants frequently discuss how their personal characteristics - including educational background, socioeconomic status, professional experience, and neurodiversity - shape their responses. Many believe their reactions deviate from the perceived norm, either describing themselves as more analytical compared to a generally more "empathetic" public, or reporting stronger emotional engagement than average. Notably, several participants challenge the very concept of a "general public," emphasizing the diversity of perspectives and questioning such generalizations, with one observing that readers of certain newspapers are "conditioned" to react with greater anger to headlines.

A.8.4 Q4

When asked about the influence of their media consumption habits (Q4), an interesting disconnect emerges. Many annotators explicitly state that their media consumption patterns do not affect their responses, yet their explanations reveal deep-seated attitudes toward different news sources. This apparent contradiction stems from annotators viewing their skepticism toward certain platforms and outlets not as a "consumption pattern" but as a fundamental approach to information processing. Many annotators express a high degree of distrust towards social media platforms such as Facebook as a primary news source and towards tabloid outlets, contrasting these with more trusted, traditional sources like the BBC. One annotator, highlighting their distrust of certain outlets, states that they avoid tabloids because they are "just nonsense really, proper biased" while another express a general suspicion of Facebook posts, viewing the platform as more for social interaction than trustworthy news. However, they do not view these preferences as biasing their responses, but rather as applying consistent critical evaluation. Additionally, annotators broadly fall into two groups regarding their approach to source evaluation. The first group reports that source credibility significantly influence their emotional engagement, with one noting they "don't get as worked up about stories from dodgy sources." The second

group emphasizes prioritizing content over source, with one explaining they "only consider the content, not the publisher."

A.8.5 Q5

Reflecting on the most salient factors shaping their emotional responses (Q5), annotators frequently emphasize an interplay of several key elements. Personal background such as upbringing and professional experience emerge as particularly important. Similar numbers of annotations mention news content, with annotators particularly responsive to stories involving injustice, vulnerable populations (especially children), and issues of immediate personal relevance. One annotator powerfully illustrate this interaction between these two factors, explaining how "growing up working class" might have instilled a certain resilience, yet emphasizing that this does not diminish their emotional response to the suffering of innocent individuals, especially children. The perceived credibility of news sources is also a factor, with one annotator articulating how "I take proper news sources more seriously." The presentation style of news content - including emotional language, imagery, and formatting - also influence responses, with several annotators demonstrating awareness of "sensationalised" content and clickbait tactics. Notably, these factors often operate interactively rather than in isolation.

A.8.6 Q6

Finally, when considering their experience with the annotation task itself (Q6), many annotators report that the task heighten their awareness of emotional responses to news. Some note that the annotation process make them more consciously aware of their emotional responses, prompting deeper reflection on the quality and factual nature of the news. As one annotator puts it, "I found myself properly thinking about how each story affected me." Participants frequently discuss their news evaluation strategies, considering multiple factors including source credibility, visual elements, and headline framing. Notably, contrary to common assumptions about social media engagement, many annotators express reluctance to share news on social platforms, with one annotator stating that they "don't share news on social media at all."

The responses reveal important individual differences in emotional engagement with news. Some annotators, particularly those identifying as neurodivergent, describe carefully managing their emotional engagement to avoid exhaustion, noting that news stories can trigger intense, lasting emotional responses. Others report preferring to reserve emotional energy for personal relationships rather than news content. Content preferences emerge as another key theme, with participants expressing greater interest in positive news, scientific developments, and locally relevant stories, while showing less engagement with celebrity news or sensationalized content.

A.8.7 Methodological Limitation

While our qualitative analysis provides valuable insights, several methodological limitations warrant discussion. First, resource constraints necessitate written questionnaires rather than in-depth interviews, potentially limiting the nuance and richness of responses. Second, the opt-in nature of the interview participation may introduce selection bias, as annotators willing to provide detailed written responses might not fully represent our broader annotator population. Third, since the interviews are conducted after both the main annotation task and the persona variable survey, participants' responses might have been influenced by these prior experiences.

These design elements ultimately strengthen rather than compromise our findings. The post-task timing of the interviews prove advantageous, allowing annotators to develop more nuanced reflections on their annotation process and emotional responses. While our participants may represent more engaged annotators, their detailed accounts provide exactly the kind of rich, experiential data needed to complement our structured persona variables. The qualitative insights thus serve their intended purpose: providing crucial context that enriches our understanding of the quantitative patterns observed in our regression analysis.

A.9 Example Image Textual Description

We show an example news post as well as the textual description generated with the prompt in A.6.

Figure 12: An example news headline.

> The image shows a small shark being held by a person wearing gloves. The shark is facing the camera and appears to be alive. The person holding the shark is slightly out of focus, with only the lower part of their face and body visible. The background is a blurry expanse of ocean. The BBC News logo is overlaid on the bottom left corner of the image. The text and image work together to tell the story of sharks testing positive for cocaine off the coast of Brazil. The image provides a visual representation of the subject matter, while the text gives context and further information. The visual element that grabs attention is the shark itself. Its size and proximity to the camera create a sense of immediacy and make the story feel more tangible. The contrast between the small, seemingly harmless shark and the serious implications of it testing positive for cocaine adds a layer of complexity to the image.

A.10 Predicting Individual Affective Arousal Prompt

For textual input:
System prompt (for the condition with persona);

```
Today is {current_date}. You are from
↪  the United Kingdom. Your first
↪  language is English. Here are how
↪  you answered a list of questions
↪  about yourself:
Question: {persona_question}
Your Answer: {persona_answer}
......
```

User prompt:

```
headline_input = A Facebook post from
↪ {page_name} posted on {post_date}:
{headline_text} The post contains an
↪ image where: {texual_description}

Article headline: {headline}

Engagement metrics:
• Emoji/reaction count:
↪ {total_interaction}
• Comments: {comments}
• Shares: {shares}

The arousal scale ranges from very calm
↪ (1) to very active (7). At the calm
↪ end of this scale (1), you feel
↪ completely relaxed, calm, sluggish,
↪ dull, sleepy, or unaroused. If you
↪ feel completely calm, indicate this
↪ by choosing 1 (very calm). At the
↪ active end of the scale (7), you are
↪ stimulated, excited, frenzied,
↪ jittery, wide-awake, or aroused. If
↪ you feel completely aroused, choose
↪ 7 (very active). If you are not at
↪ all excited nor at all calm, choose
↪ 4 (neutral). Choose in-between
↪ options to indicate intermediate
↪ levels of excitement or calmness.

How calm vs. active do you feel after
↪ reading this news headline?
↪ (Arousal)
1 (very calm)
2 (calm)
3 (somewhat calm)
4 (neutral)
5 (somewhat active)
6 (active)
7 (very active)
Please respond with a single number from
↪ 1-7.

If few-shot:
    messages.append("role": "user",
    ↪ "content": {headline_input})
    messages.append({"role":
    ↪ "assistant", "content": {label})
messages.append({"role": "user",
↪ "content": {headline_input})
```

For image-input condition, the prompt is the same except that the post-specific textual description is replaced by the news post screenshot.

A.11 Additional Few-shot learning results

We present additional few-shot results in Table 9 and Figure 13.

Modality	Setting	Performance Metrics		
		MAE ↓	Accuracy ↑	Within±1 Accuracy ↑
Text	0-shot, no persona	1.035	29.4	74.4
	0-shot, with persona	**0.914**	**36.4**	**78.8**
	4-shot, no persona	1.029	31.6	74.1
	4-shot, with persona	1.016	32.1	75.5
	8-shot, no persona	0.955	35.1	77.7
	8-shot, with persona	0.971	36.1	76.0
	16-shot, no persona	0.851	39.7	81.5
	16-shot, with persona	0.824	42.3	82.0
	32-shot, no persona	0.812	42.1	83.4
	32-shot, with persona	0.782	44.4	83.6
Image	0-shot, no persona	0.936	37.0	77.0
	0-shot, with persona	**0.841**	**39.6**	**82.0**
	4-shot, no persona	1.116	29.9	71.3
	4-shot, with persona	1.069	30.6	73.1
	8-shot, no persona	1.133	27.8	72.2
	8-shot, with persona	1.074	29.4	74.6
	16-shot, no persona	1.054	33.3	73.9
	16-shot, with persona	0.959	35.9	77.9
	32-shot, no persona	0.926	39.2	77.5
	32-shot, with persona	0.858	42.8	79.6

Table 9: Performance comparison of text-based and image-based models across few-shot settings and persona conditions. Lower MAE (↓) and higher Accuracy and Within±1 Accuracy(↑) indicate better performance. Best zero-shot results are in **bold**, and best overall results per modality are underlined. Accuracy and Within±1 accuracy are shown as percentages.

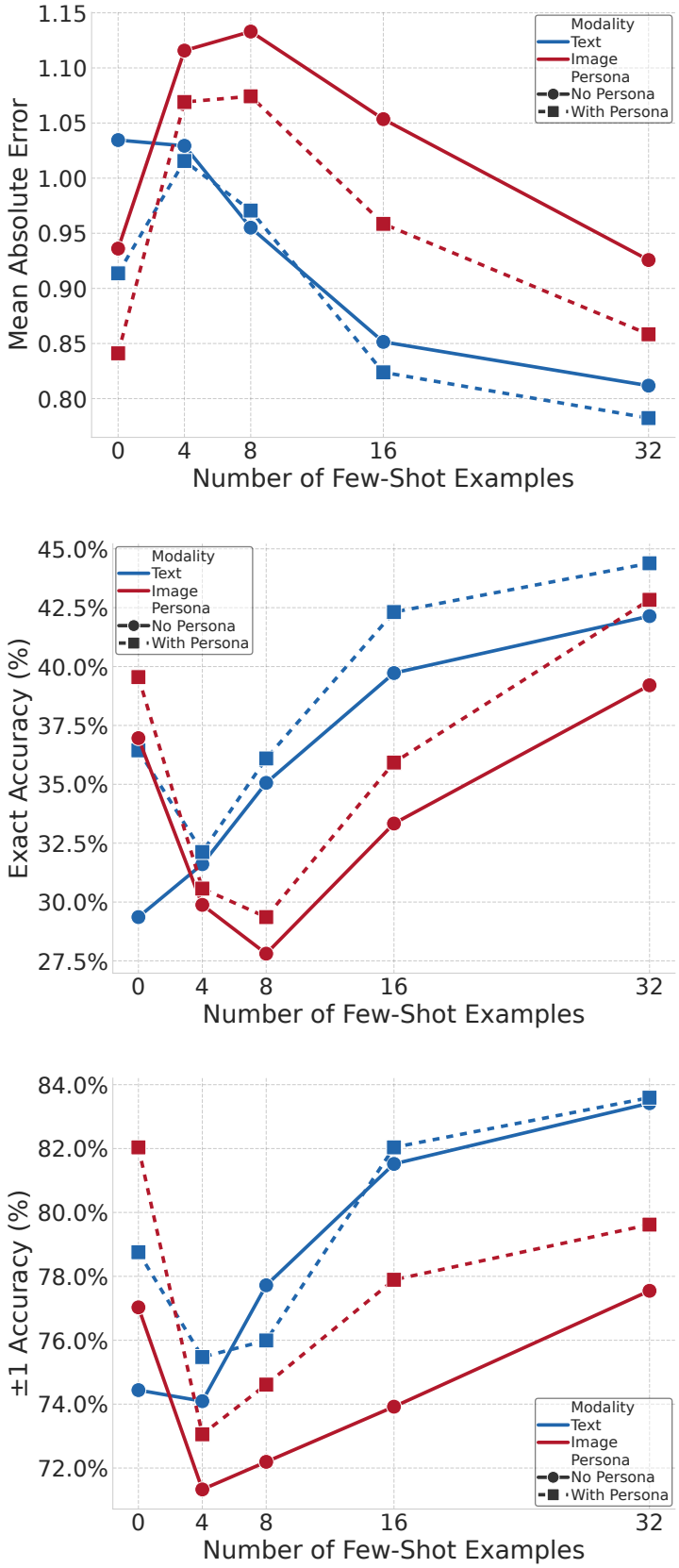


Figure 13: Few-shot learning performance, measured by MAE, exact match accuracy (%), and ±1 accuracy (%), as a function of the number of few-shot examples (0, 4, 8, 16, and 32).