

# Supplementary Material

## Table of Contents

<b>Part 1 - The PHORECAST Dataset</b>	
• A. Dataset Access	13
• B. Recruitment Process	13
• C. Curation Rationale	13
• D. Integration with existing public health communication frameworks	13
• E. Ethical Consideration	13
• F. Terms of Use	14
• G. Data Rights Compliance and Issue Reporting	14
• H. Informed Consent	14
• I. Survey Details	15
• J. Data Processing	18
• K. Data Analysis	18
• L. Media Analysis	18
<b>Part 2 - Training Details and Evaluation Metrics</b>	
• L. Training Details	22
• M. Dataset Preparation	22
• N. Free Form Metrics	23
• O. Generalization	23
• P. Qualitative Examples	24
<b>Part 3 - Generalizing Response Predictions to Unseen Communication Strategies</b>	32
<b>Part 4 - Future Practical Use Case: Communication Strategy Recommendation</b>	36

## A THE PHORECAST DATASET

### A.1 ACCESS

The dataset and scripts will be released upon acceptance.

### A.2 RECRUITMENT PROCESS

We recruit participants through the <https://www.surveycircle.com/en/> platform and via the social media platform LinkedIn. Each participant is incentivized with a \$10 Tango gift card. Participants are provided with details of potential risks of participating in the survey and their ability to voluntarily skip questions or end the survey before completion. All data was collected between Jan 5th, 2025 to Jan 9th, 2025.

### A.3 CURATION RATIONALE

The PHORECAST dataset aims to map real human profiles (demographics, personality, and locus of control) to their responses / reactions from interacting with various public health campaigns. The primary purpose is for academic research to study how different people interact with stimuli and simulate how and why different communities respond differently to visuals. The results will be used to build an AI simulator that can mimic real world communities.

### A.4 INTEGRATION WITH EXISTING PUBLIC HEALTH COMMUNICATION FRAMEWORKS:

Our approach is designed to complement existing public health communication frameworks by offering a scalable, low-risk method for pre-testing messages prior to real-world deployment. Currently, message evaluation relies heavily on human subjects through focus groups or randomized trials — ad-hoc approaches that are often resource-intensive, time-consuming, and difficult to implement in time-sensitive or emergency contexts.

By simulating audience responses using LLM-generated communities, our method offers practitioners a novel way to assess potential message effectiveness and refine content in advance. This novel approach for public-health messaging has the potential to reduce reliance on real-time, post-dissemination adjustments, which can be both costly and disruptive, potentially biased and less representative given the limited sampled community groups and resource-constraints.

In practice, this new approach, empowered by LLM and broader online community surveys, supports:

1. Rapid iteration and scenario testing within existing campaign development workflows
2. Enhanced training for public health students and practitioners, analogous to surgical simulators in medicine, allowing them to practice message design and delivery in a controlled, harm-free environment
3. Improved trust and efficiency in community engagement, by minimizing the need for repeated in-person testing, particularly with vulnerable populations

While not a replacement for field validation, this newly designed AI tool aims to serve as an intermediary step that enhances message development and helps operationalize core principles from established frameworks, such as the Health Belief Model, Social Cognitive Theory, and the Theory of Planned Behavior.

### A.5 ETHICAL CONSIDERATIONS

This study is approved by the Institutional Review Board (IRB), under an exemption category. All participants provide digital informed consent prior to beginning the survey. The consent form clearly outlines potential risks and benefits of participation, and informs participants of their right to withdraw at any time and to decline to answer any questions. To maintain anonymity and separate responses from personal identifying information, participants are directed to a separate form for incentive compensation. All collected data are de-identified before analysis. We follow standard practices for data collection (Paullada et al. (2021), Quy et al. (2022)).

## A.6 TERMS OF USE

**Purpose** The dataset can only be used for educational and research purposes or to develop and evaluate AI models.

**Restrictions** This dataset is a public-use dataset as defined by the Data Procedures Manual by NCES (<https://nces.ed.gov/>). All individually identifiable information has been removed to protect the confidentiality of participants and no license is needed to access the dataset.

**Deanonymization** The users are prohibited from de-anonymizing the individuals represented in the dataset.

**Content Warning** The dataset may include text, images and videos that could be considered unsafe or offensive for some individuals. The users must use appropriate measures to filter content when used for educational or training purposes to adhere to the ethical and safety standards.

**Endorsement and Liability** The authors of the paper, the dataset creators, funders and the affiliated institution do not endorse the views and opinions expressed in the data and are not liable to damages resulting from the use of the dataset.

## A.7 DATA RIGHTS COMPLIANCE AND ISSUE REPORTING

We are committed to complying with data protection rights. If any individual whose data is included in the PHORECAST dataset wishes to have their data removed, we provide a straightforward process for issue reporting and resolution. Concerned parties are encouraged to contact the authors directly via making a formal issue reporting on our GitHub page. Upon receiving a request, we will engage with the individual to verify their identity and proceed to remove the relevant entries from the dataset. We commit to addressing and resolving such requests within 30 days of verification.

## A.8 INFORMED CONSENT

**Eligibility** Thank you for expressing interest in helping us build an AI Community Simulator. Before we get started, please answer the following eligibility questions:

Are you 18 years of age or older? ☐ Yes ☐ No

Do you currently reside in the United States? ☐ Yes ☐ No

If “no” to any one of these questions: Thank you for taking the time to express interest in our study. For more information about the work we do, please visit: [our public health engagement link] If eligible, the participants are directed to the informed consent:

**Purpose of the Study** This research project aims to develop an AI-powered community simulator modeled after real-world communities to enhance public health training and practice. This study seeks to understand how individuals’ demographic identities and personality traits interact and react to public health messages. The results will be used to develop a prototype of an AI-powered community simulator to test public health trainees’ development of public health messages.

**Procedures** You will be asked to answer questions about your social and physical identity and personality traits. Then, you will be asked to review and react to various public health communication messages. The survey will take approximately 30 minutes to complete. Once you complete the survey, you will be asked to enter your contact information separately to receive your compensation.

**Potential Risks and Discomforts** There are minimal risks or inconveniences from participating in this research study. The length of time required to take this survey (30 minutes) may be inconvenient for some; however the survey has been designed to limit this possibility. If you feel uncomfortable answering some questions, you have the right to skip any questions you do not want to answer.

**Potential Benefits** While there are no direct benefits to you for participating in this study, your involvement will significantly benefit the development of training opportunities for public health trainees to increase the effectiveness of public health messaging to improve health outcomes.

**Confidentiality** Any potential loss of confidentiality will be minimized by storing data in a password-protected computer. If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of [Our University] or governmental authorities if you or someone else is in danger

or if we are required to do so by law. You will be taken to a separate form to enter your contact information for compensation to ensure that your personal information will not be linked to your responses.

**Compensation** By participating in this study, you will receive a \$10 gift card. You will be responsible for any taxes assessed on the compensation. A separate email containing your compensation will be sent to you within 30 days of completing the survey

**Right to Withdraw and Questions** Your participation in this study is completely voluntary. You may choose not to take part at all. If you decide to participate in this study, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify. If you decide to stop taking part in the study, have any questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigators: [Removed to comply with double blind]

**Participant Rights** If you have questions about your rights as a research participant or wish to report a research-related injury, please contact: [our institutional review board] For more information regarding participant rights, please visit: [our institutional review board] This research has been reviewed according to the IRB procedures for research involving human subjects.

**Statement of Consent** Your signature indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. You will receive a copy of this signed consent form. By checking the box below, you indicate that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. If you agree to participate, please check the box: ☐ I agree

## A.9 SURVEY

We use the Big Five Inventory (BFI-2) by Soto & John (2017) along with their scoring methods<sup>1</sup>. We collect demographics information and personality traits by having participants self-report and answer the questions. Participants answer baseline questions prior to viewing any campaigns on five different health topics. We show an example of the questions asked for the topic of Chronic Diseases and Substance Use in Fig 5. Subsequently, participants engaged with five distinct campaigns related to the same baseline topics, and provided their opinions and reactions through a series of questions. We show an example of this in Fig 6.

## A.10 CAMPAIGN SELECTION PROCESS

To build the campaign repository, we curated open-source health communication materials from nonprofit organizations, peer-reviewed publications, government agencies, and other entities focused on public health behavior change. Inclusion criteria required that messages (1) addressed a health behavior or outcome and (2) included a visual component (e.g., print or video campaigns). We intentionally sampled a diverse range of campaigns to capture messages targeting different age groups and demographic segments.

**Annotation Process:** Each campaign was annotated along three dimensions:

*Target Population:* Classified into predefined age-based categories (Children  $\leq 11$ , Adolescents 12–17, Young Adults 18–24, Adults 25–44, Adults 45–64, and Older Adults  $\geq 65$ ). When messages applied to multiple groups, overlapping categories were selected.

*Message Type:* Based on the dominant communication strategy—informative, persuasive-efficacy, or persuasive-threat—categories commonly used in public health communication and grounded in established health behavior theories.

*Health Behavior & Outcome:* Specific health behaviors and linked outcomes were identified for each message.

<sup>1</sup>For more information, please visit <http://www.colby.edu/psych/personality-lab/>



Please answer the following questions on a scale of 1 (not at all) to 9 (extremely)

A score of 5 indicates a neutral opinion.

	1	2	3	4	5	6	7	8	9
I am concerned about the health risks of chronic diseases	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am motivated to not ignore chronic disease symptoms	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ignoring chronic disease symptoms is harmful to my health	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am open to regular health screenings and preventive care in the future	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please answer the following questions on a scale of 1 (not at all) to 9 (extremely)

A score of 5 indicates a neutral opinion.

	1	2	3	4	5	6	7	8	9
I am concerned about the health risks of using substances	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am motivated to not use substances	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Substance use is harmful to my health	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am open to trying a substance in the future	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 5: Example of baseline opinions we ask participants prior to viewing any public health campaigns. Each participants answers four baseline questions on five health topics, such as Chronic Diseases, Substance Use, Smoking/COPD, Nutrition etc

# CARDIOVASCULAR DISEASE

## THE WORLD'S NUMBER 1 KILLER

Cardiovascular diseases are a group of disorders of the heart and blood vessels, commonly referred to as **heart disease** and **stroke**.

20.5 MILLION

deaths every year from CVD

33%

of all global deaths

>75%

of CVD deaths take place in low- and middle-income countries

### GLOBAL CAUSES OF DEATH

### RISK FACTORS FOR CVD

- High Blood Pressure
- High Cholesterol
- Overweight & Obesity
- Air Pollution
- Physical inactivity
- Unhealthy Diet
- Diabetes
- Tobacco
- Kidney Disease
- Harmful use of alcohol

Sources: World Health Organization; IHME, Global Burden of Disease

info@worldheart.org  
www.worldheart.org

worldheartfederation  
worldheartfed  
worldheartfederation

Type in every thought that came to mind viewing this material

To what extent did the material make you feel:

Scale: 1 (not at all) -9 (extremely)

1 2 3 4 5 6 7 8 9

A score of 5 indicates a neutral opinion.

Sad

Angry

Afraid

Guilty

Disgusted

Worried

Ashamed

Hopeful

The message makes me more concerned about the **health risks of heart disease**.

1 2 3 4 5 6 7 8 9

A score of 5 indicates a neutral opinion.

Scale: 1 (not at all) - 9 (extremely)

The message motivates me to **not ignore cardiovascular health**

1 2 3 4 5 6 7 8 9

A score of 5 indicates a neutral opinion.

Scale: 1 (not at all) - 9 (extremely)

In your opinion, how **harmful** is **ignoring heart health** to your general health?

0 1 2 3 4 5 6

A score of 3 indicates a neutral opinion.

Scale: 0 (not at all)-6 (extremely harmful)

How open are you to **maintaining heart-healthy practices** in the future?

1 2 3 4 5 6 7 8 9

A score of 5 indicates a neutral opinion.

Scale: 1 (not at all)-9 (extremely)

Figure 6: An example of the public health campaign survey shown to participants. This snippet illustrates the diverse question types, including free-form responses, 4-point Likert scales (gauging concern, motivation, harm perception, and openness), and emotional assessments. Note: Display order here does not reflect the actual survey flow.

The initial coding was conducted by trained research assistants under the guidance of the principal investigators. All annotations underwent a secondary review by the investigators to ensure coding consistency. Any discrepancies were resolved through discussion until consensus was reached.

#### A.11 DATA PROCESSING

Due to a discrepancy between compensation forms and completed surveys, we implemented several measures to ensure data integrity. Incomplete surveys and submissions from duplicate IP addresses were removed. Additionally, we identified and excluded four participants whose free-form responses exhibited characteristics consistent with generative model output, indicating potential fabrication.

#### A.12 DATA ANALYSIS

We look into the covariance matrix of our features to analyze their relationships. We observe high correlation between different traits such as Respectfulness and Agreeableness, Extraversion and Energy Level, Negative Emotionality, Depression and Emotional Volatility. Please refer to Fig 7 for the full analysis. We also analyze the correlation between demographics or personality and the responses. We observe that the locus of control and the sociability level of the individual greatly impact their emotional responses (Fig 8). Furthermore, an individual's race/ethnicity greatly impacts their emotional response, followed by their profession and education level. Those features also impact the harm perception, concern level, motivation and openness levels.

We analyze the distributional differences in responses among different groups including gender (Fig 10), political affiliation (Fig 11), and education levels (Fig 12). We make some key observations:

Women and men generally exhibit similar opinions prior to interacting with any campaigns. Post-intervention, women report slighter higher levels of concern across most categories. Notably, sexual health interventions appeared to elicit the strongest responses among non-binary participants.

Our dataset further indicates significant variations in how different political affiliations perceive and respond to health-related topics. For instance, while conservatives exhibit high initial concern regarding mental health and nutrition, independents demonstrate increased concern scores following the stimulus. A unique pattern emerges among libertarians, who show the highest openness to smoking, yet the lowest propensity for vaccination or dietary care. Conversely, both libertarians and independents demonstrate the highest levels of concern and motivation regarding timely vaccination.

Analysis of health-related behaviors and perceptions by education level reveal that individuals with higher educational attainment consistently report elevated concern for various health topics. Notably, for all categories except nutrition, the disparity in concern across education levels diminishes post-intervention, suggesting potential for more effective health interventions among lower education groups.

As illustrated in Figure 9 participants with a high internal LOC consistently exhibit elevated levels of concern, motivation, harm perception, and openness, both pre- and post-stimulus, across all health topics. This finding aligns with the theoretical framework of internal LOC, where individuals' belief in their personal control over outcomes naturally correlates with greater concern, motivation, and risk perception. Conversely, individuals with a high external LOC, who attribute outcomes largely to external factors, tended to display comparatively lower personal concern and motivation.

#### A.13 MEDIA ANALYSIS

We analyze the effectiveness of different campaign types by (1) analyzing what campaigns are most effective for different health topics, (2) which elicited the highest levels of emotion overall and (3) we split participants with low and high big 5 personality traits and analyze the differences in emotion scores. As seen in Fig 13 different personality traits react and perceive different types of campaigns differently. For example, people with high levels of anxiety tend to feel more afraid, angry, ashamed, guilty and worried when viewing persuasive images as opposed to threatening or informational. On the contrary, people with low anxiety tend to react more strongly to threatening campaigns. We analyze the effect of the media types across different categories of locus of control in Fig 14 and by race/ethnicity in Fig 16. Finally, we hope this work inspires researchers to study the aesthetic or psychological features in the pixel space that impact how different individuals react and respond to marketing content.

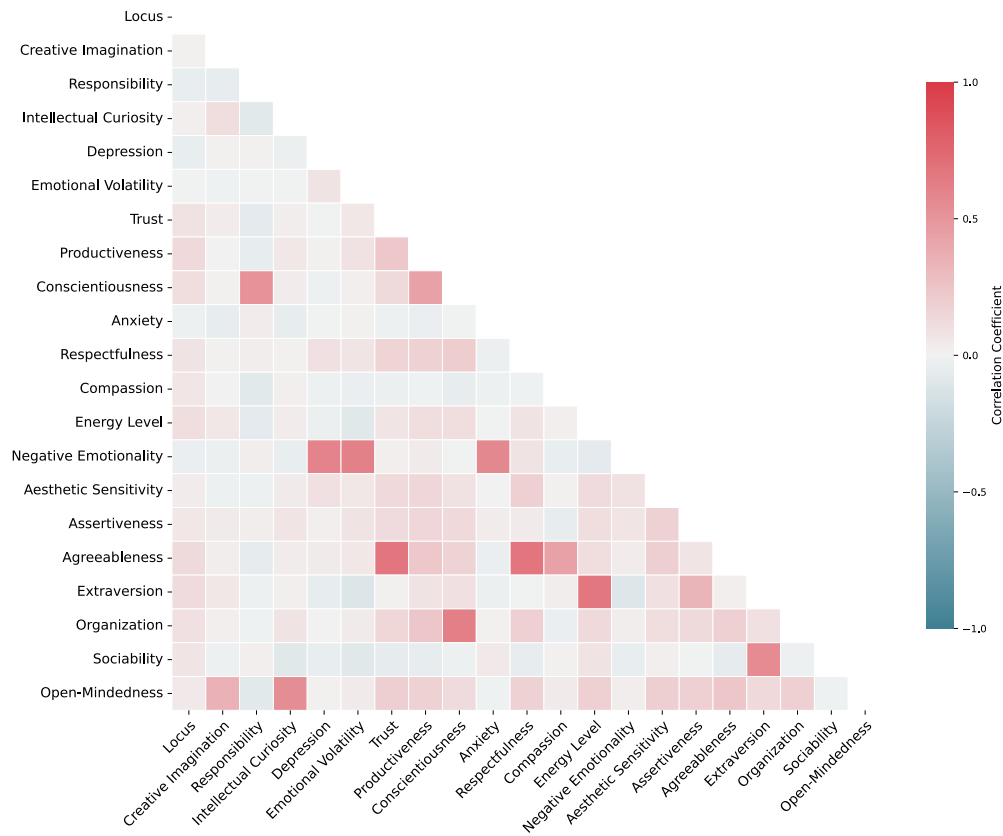


Figure 7: Covariance matrix of personality features. Agreeableness, Trust, Productiveness, Respectfulness and Compassion are highly correlated. Emotional Volatility, Anxiety, Depression and Negative Emotionality are highly correlated. The Locus of Control is more highly correlated with Productiveness, Agreeableness, Extraversion, Organization, Energy Level than it is with Creative Imagination, Responsibility, Depression and Emotional Volatility.

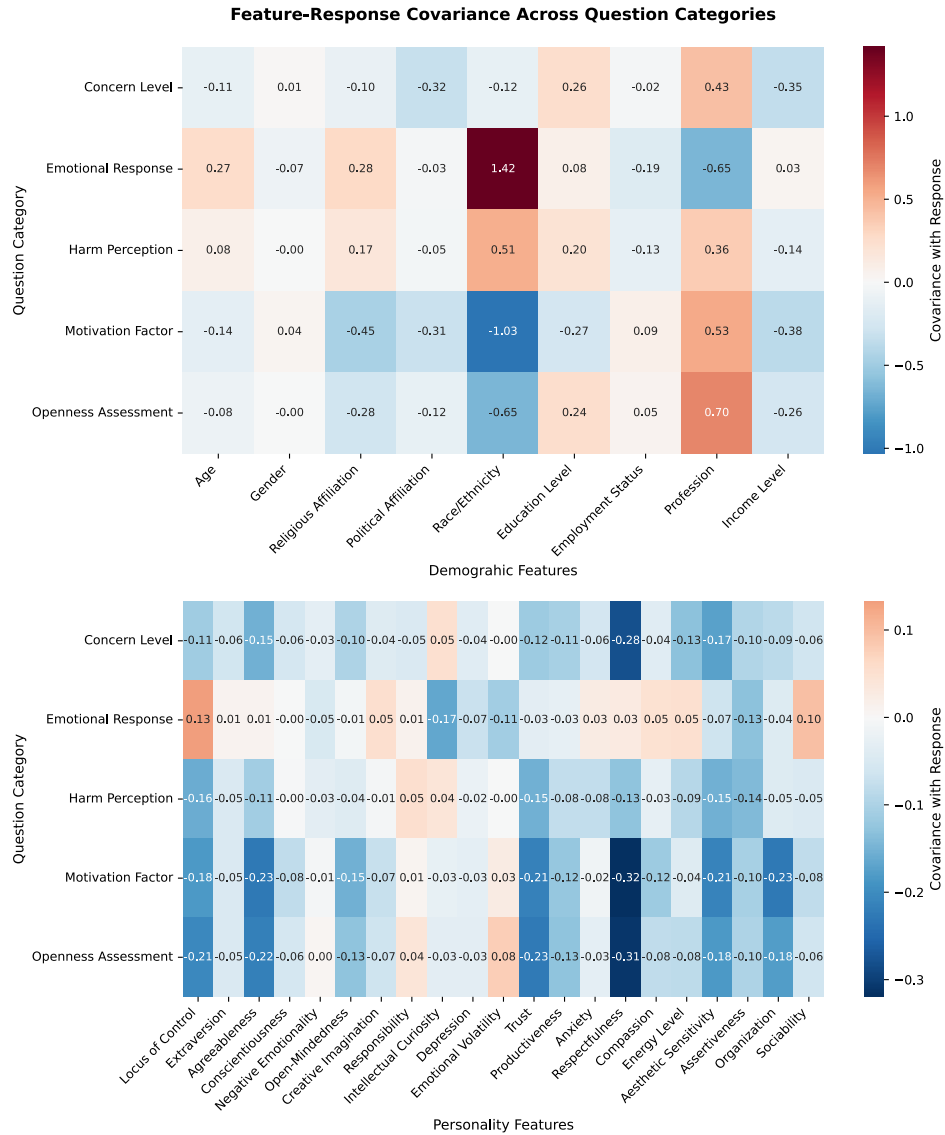


Figure 8: Demographics & Personality Features and Response covariance across question categories. Race/Ethnicity shows strong correlations with emotional responses (cov = 1.42) and harm perception (cov = 0.51), but minimal influence on motivation or openness. Profession is a dominant factor across multiple dimensions, showing high covariance with openness (cov = 0.70), motivation (cov = 0.53), concern level (cov = 0.43), and harm perception. Within personality traits, emotional volatility and sociability are most strongly associated with emotional responses and openness. Locus of control exhibits the highest covariance among personality traits with emotional responses (cov = 0.13)

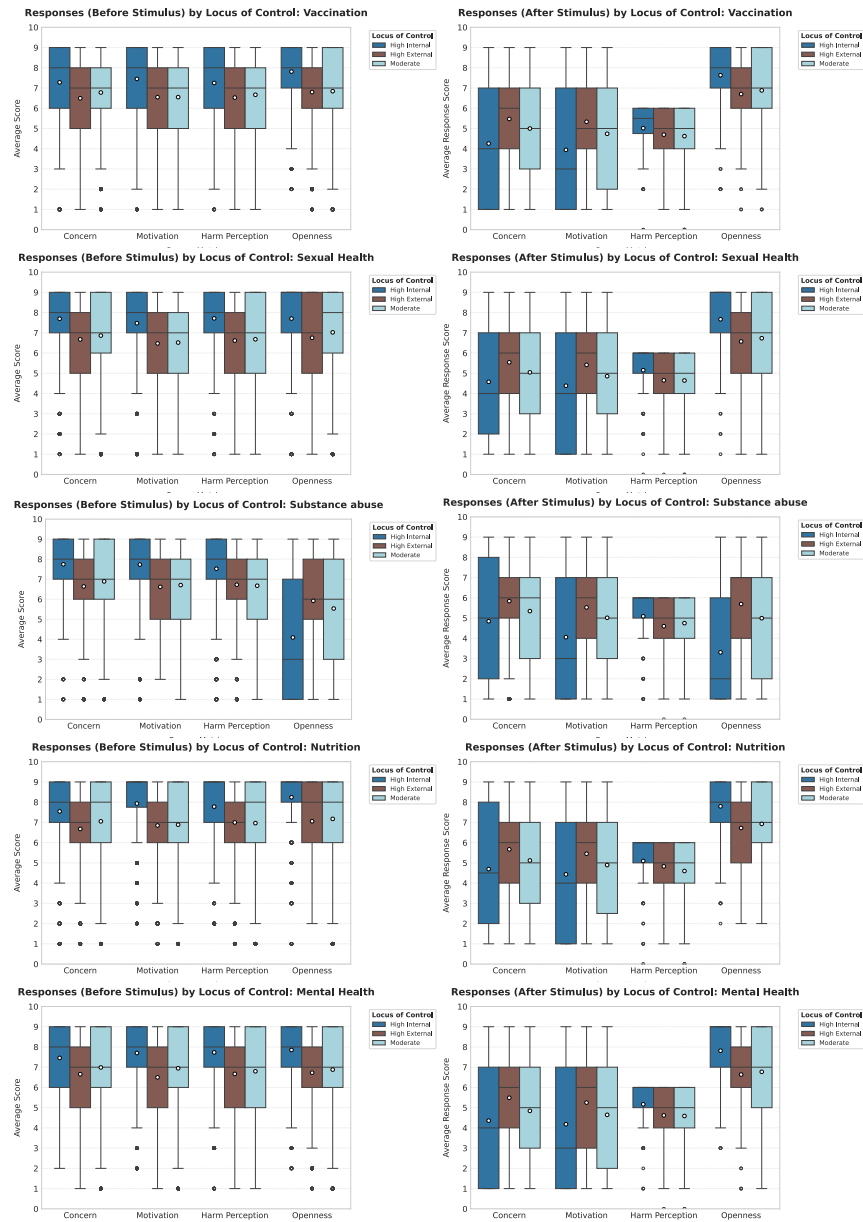


Figure 9: Concern, Motivation, Harm Perception and Openness of individuals with different Locus of Control prior to viewing any marketing content. People with a *high internal* locus are more concerned about the risks of health concerns and motivated to practice behaviors to promote their health. This group is less open than others to abuse substances or smoke. People with a *high external* locus are more open to smoking and abuse substances than people with a *moderate* locus.



## B TRAINING DETAILS AND EVALUATION METRICS

### B.1 TRAINING DETAILS

We use the unsloth [Daniel Han & team \(2023\)](#) framework for training the models with the following hyperparameters:  $bs = 1$ ,  $lr = 2e - 4$ , and rank  $r = 8$  for Lora [Hu et al. \(2021\)](#). We use base models *unsloth/Llama-3.2-11B-Vision-Instruct* and *unsloth/gemma-3-12b-it*. We utilize *paged\_adamw\_8bit* optimizer and a max sequence length of 2048. We always keep the vision layers frozen, and only tune the attention, language and MLP layers. We train for a total of 12k steps.

### B.2 DATASET PREPARATION

The training data comprises each participant’s responses to five health advertisements. This is randomly chosen for each participant at the start of the survey. Input to the model is randomized by probabilistically including features: demographics/personality (90% of the time, for which we then choose a random number of features from 1 to  $\text{len}(\text{avail\_demographics})$ ), locus of control (75%), free-form text (30%), and baseline/in-context opinions (50%, for which we choose a random number of features to include). This randomization method is used to create the training and validation split for model evaluation, as shown in Table 8. After training, we evaluate the model on new advertisements and new individuals. As described in 4, we construct our validation set by ensuring our sampling includes unique representatives for each gender, religion and race/ethnicity.

The model receives a task description, along with participant demographics, personality traits, and in-context question/answer pairs. Fig. 15 illustrates our chat template when all available features are utilized, a scenario that can occur within our feature randomization process.

Attribute	Train Count	Train %	Val Count	Val %
<i>Number of Individuals</i>	1010	–	521	–
<i>Total Samples</i>	27,572	–	8,537	–
<i>Gender</i>				
Male	15,596	56.6%	4,422	51.8%
Female	11,748	42.6%	4,050	47.4%
Other	228	0.8%	65	0.8%
<i>Religion</i>				
Christian	20,079	72.8%	6,432	75.3%
Unknown	4,087	14.8%	1,014	11.9%
Muslim	1,833	6.6%	663	7.8%
Other	1,573	5.7%	428	5.0%
<i>Race</i>				
White	17,409	63.1%	5,568	65.2%
Black	8,019	29.1%	2,215	25.9%
Hispanic	1,039	3.8%	273	3.2%
Asian	611	2.2%	182	2.1%
Indigenous	481	1.7%	273	3.2%
<i>Media Count</i>				
Unique Media	29	–	8	–

Table 8: Demographic distribution and media counts across training and validation splits. Percentages are computed within each split. We keep one unique image per topic to validate the model’s ability to generalize to unseen images. We ensure we have representative individuals based on three demographical features: gender, religion, and race. This experimental design creates a training split of 1010 unique individuals with 27, 572 unique samples, and a validation split of 521 individuals (42 unique individuals) with their corresponding 8537 samples. Our dataset is dominated by White, Christian males.

### B.3 ADDITIONAL METRICS FOR FREE FORM EVALUATION

In addition to semantic similarity, we utilize the following metrics to analyze LLM generated free form responses.

**SDE Score** In keeping with evaluation standards for generative images, we compute the statistical distribution of embedding (SDE) features –stratified by personality traits to assess distributional alignment across different subgroups, i.e. between distributions of real and machine-generated responses, stratified by personality trait bins. SDE quantifies the similarity (lower is better) between machine and human responses by embedding them into a feature space, fitting a multivariate Gaussian distribution to each, and measuring the distributional overlap. Let the set of embeddings from human (ground truth) responses be characterized by mean vector  $\mu_r$  and covariance matrix  $\Sigma_r$ , and let the embeddings of machine-generated responses be described by mean vector  $\mu_g$  and covariance matrix  $\Sigma_g$ . The SDE is defined as:

$$\text{SDE} = \|\mu_r - \mu_g\|^2 + \text{Tr} \left( \Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right)$$

A lower SDE indicates that the generated responses more closely resemble the statistical distribution of real human responses in the embedding space. We compute SDE scores separately for subsets of data grouped by trait levels (e.g., low vs. high agreeableness). This allows us to measure the model’s fidelity in reproducing nuanced, trait-dependent response patterns. A monotonic decrease in SDE over training steps signals improved generation quality and diversity, and a low trait-specific SDE suggests the model is capturing the distinctive structure of human responses conditioned on that trait—serving as a principled metric for evaluating personality-conditioned generative alignment. We show SDE scores for Llama and Gemma across checkpoints in Fig 17

**Perplexity** We employ ‘perplexity’, a measure of uncertainty in predicting the next word in a sequence, as a discriminative metric to evaluate the model  $\mathcal{M}$ ’s ability to capture individual-specific language patterns and semantic styles. For any distinct pair of individuals, A and B, with profiles  $P_A, P_B$  and observed responses  $R_A, R_B$ , we compute the full sequence perplexity  $\text{PPL}(P_X \circ \text{"Response: "} \circ R_Y)$  for  $X, Y \in \{A, B\}$ , where  $\circ$  denotes string concatenation. For each response  $R_Y$ , a correct attribution is defined if  $\text{PPL}(P_Y \circ \text{"Response: "} \circ R_Y) < \text{PPL}(P_{X \neq Y} \circ \text{"Response: "} \circ R_Y)$ . This evaluation measures  $\mathcal{M}$ ’s capacity to recognize and associate responses with their generator’s style.

Model	Perplexity ( $\uparrow$ )	SDE ( $\downarrow$ )
<b>Llama 11B</b>		
Initial checkpoint (1k steps)	46.20	1.02
After training	<b>51.67</b>	<b>0.67</b>
<b>Gemma 12B</b>		
Initial checkpoint (1k steps)	50.76	1.04
After training	51.52	0.74

Table 9: Comparison of human language emulation performance before (1k-step checkpoint) and after training. **Metrics:** Higher accuracy (% correct) and lower Statistical Distribution of Embeddings (SDE) indicate better alignment with human responses. **Key findings:** (1) Llama shows improvement in SDE (1.02  $\rightarrow$  0.67) and an increase in accuracy (of over 5%). (2) Gemma also achieves some gains: accuracy rises from 50.76 to 51.52, and SDE improves from 1.04  $\rightarrow$  0.74. **Qualitative:** Base models often default to uncertain or repetitive language (e.g., Llama’s “I’m not sure...” in 80% of cases).

## C GENERALIZATION

We use our trained LLama model to investigate how robust we are to distributional shifts. We demonstrate strong generalization to unseen subpopulations in 10 This evaluation setting tests a

difficult case, where each sample in the validation set contains a masked number of features. For instance, a model might need to predict the participant’s response based solely on that person’s locus of control. We use the same practice (methods and hyperparameters) described in the paper but we train the models for approximately 8k steps and test on a random sample of 1k. We report accuracy with  $\pm 2$  tolerance.

Test Set	Llama (Before Training)	Llama (After Training)
Females aged 25–34	56.2	72.9
Individuals in technology making \$100k+	51.0	72.7
Non-Christian men	50.7	76.6

Table 10: Performance of Llama before and after training on different demographic test sets.

We also investigate model performance on unrelated benchmarks after training on our dataset, to test whether general language modeling capabilities degrade as a result of finetuning. We use the `lm-evaluation-harness` and evaluate Llama 3.2 11B Instruct and Gemma 3 12B models (batch size = 8) before and after training with PHORECAST (Table 11).

Task	Gemma (Before → After)	Llama (Before → After)
TriviaQA	27.6 → 46.5	51.4 → 39.7
ToxiGen	56.8 → 58.8	53.8 → 56.8
HellaSwag	62.7 → 61.1	59.2 → 57.9
MMLU	71.5 → 68.6	68.0 → 61.6

Table 11: Generalization performance of Gemma and Llama before and after PHORECAST finetuning.

While performance remains stable overall for toxigen, hellaswag, and mmlu, we observe that training with PHORECAST did lead to major changes in triviaqa scores, with Gemma improving substantially and Llama degrading.

## D QUALITATIVE EXAMPLES

We compare model responses before and after training, as illustrated in Figure 18. Prior to training, the Llama model frequently struggles with task comprehension, often prioritizing the interpretation of visual tokens over emulating the described individual. Post-training, model responses more closely resemble the true human responses. Furthermore, trained models demonstrate a notable ability to generalize to individuals who did not respond, accurately predicting an ‘none’ output. This intriguing capability motivates further investigation into theories of fear appeal effectiveness. Our future work will focus on identifying fear-based and danger-based responses, as well as refining the classification of non-responses, informed by these theoretical frameworks.

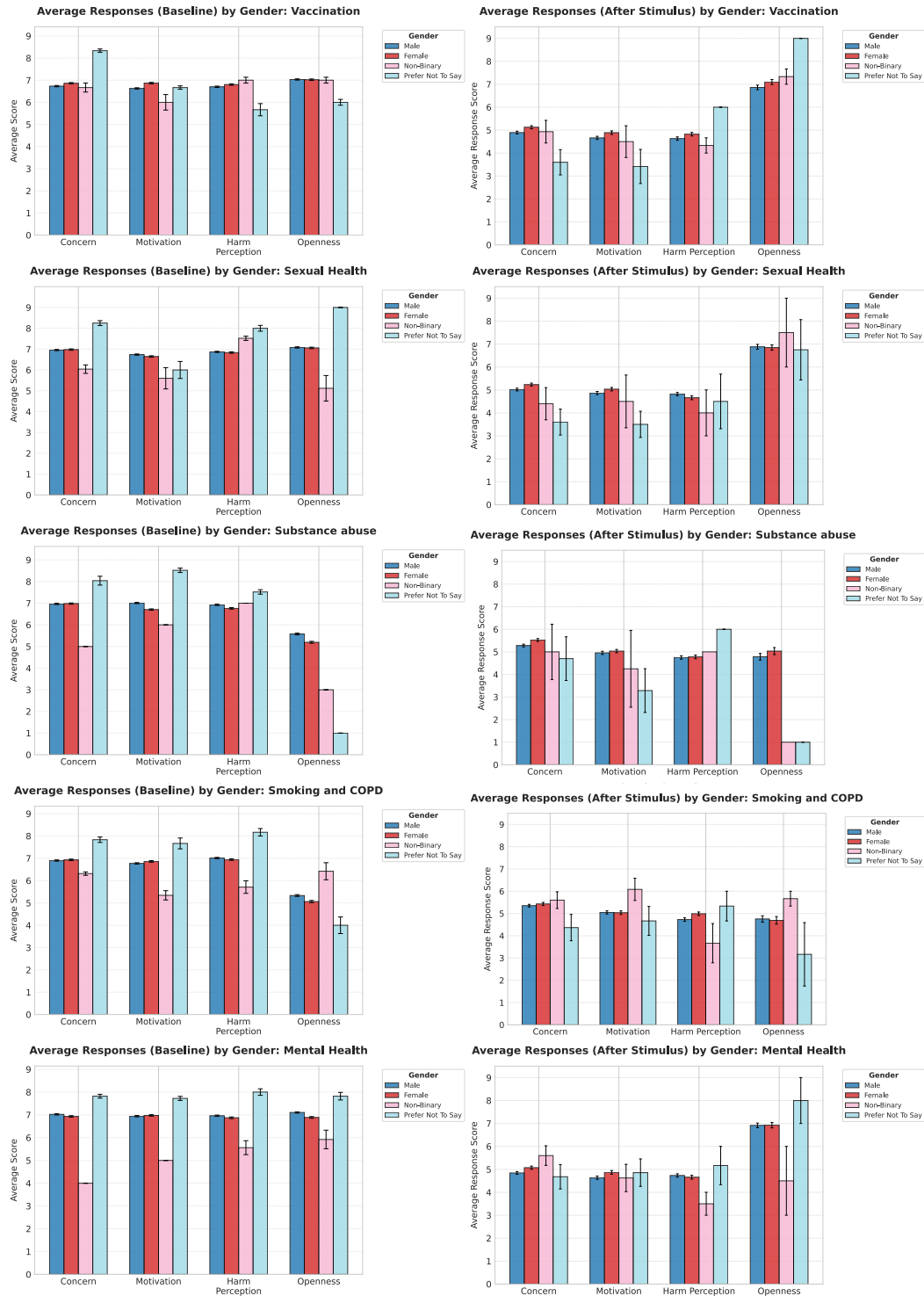


Figure 10: **Average Responses by Gender:** While baseline levels of concern, motivation, harm perception, and openness were generally similar between men and women, women exhibited a greater change in responses following exposure to marketing stimuli. Non-binary and other genders show elevated perception of sexual health and HIV risks.

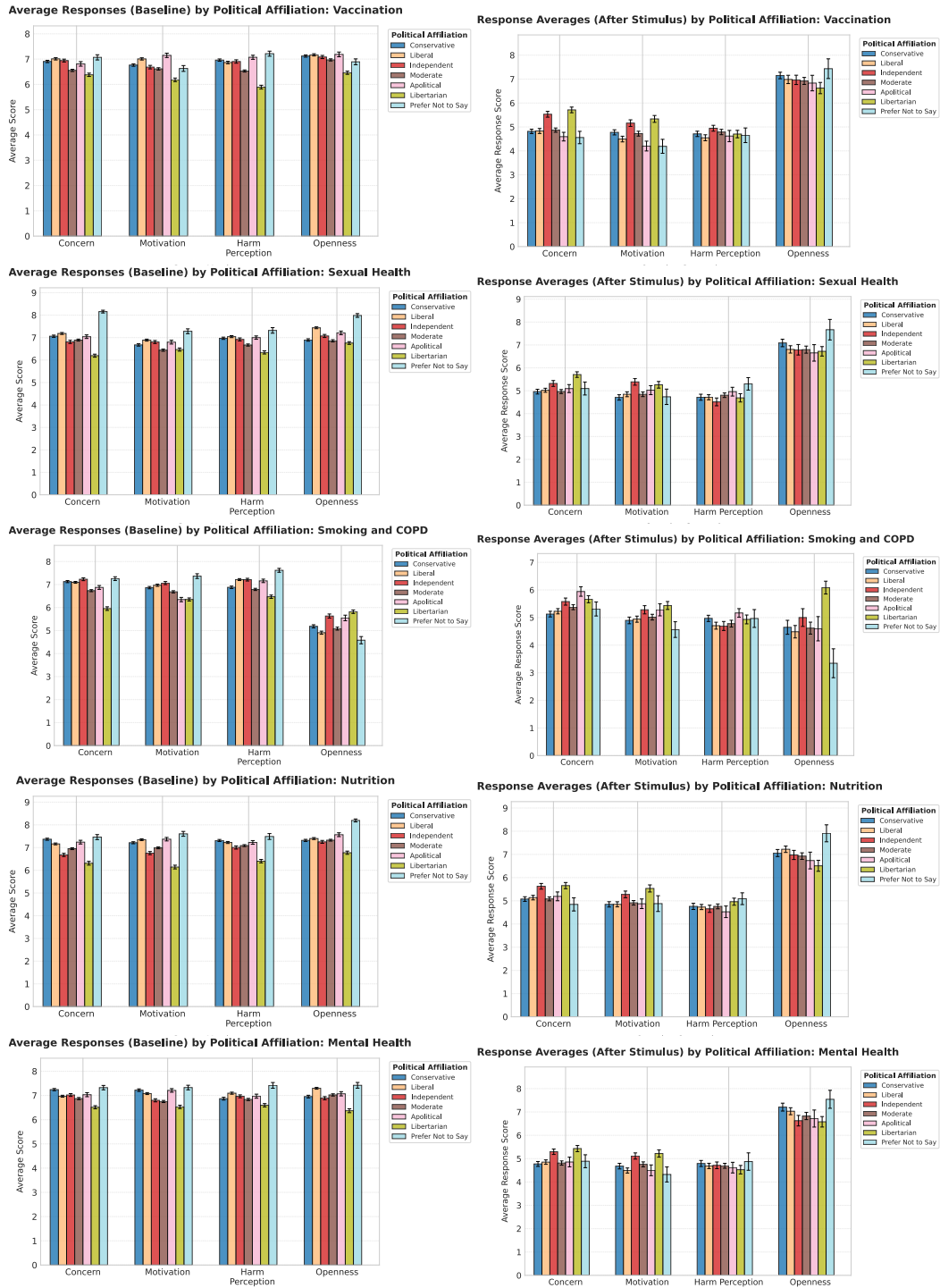


Figure 11: **Average Responses By Political Affiliation:** We analyze the opinions of different affiliated groups. We observe that Libertarians appear to have the least amount of concern, motivation and harm perception regarding various health topics. At the same time, the campaigns seem to have a bigger effect on them. Yet, they seem to be the most open to smoke. Conservatives seem to be less concerned about the risks of skipping vaccination, mental health and nutrition.

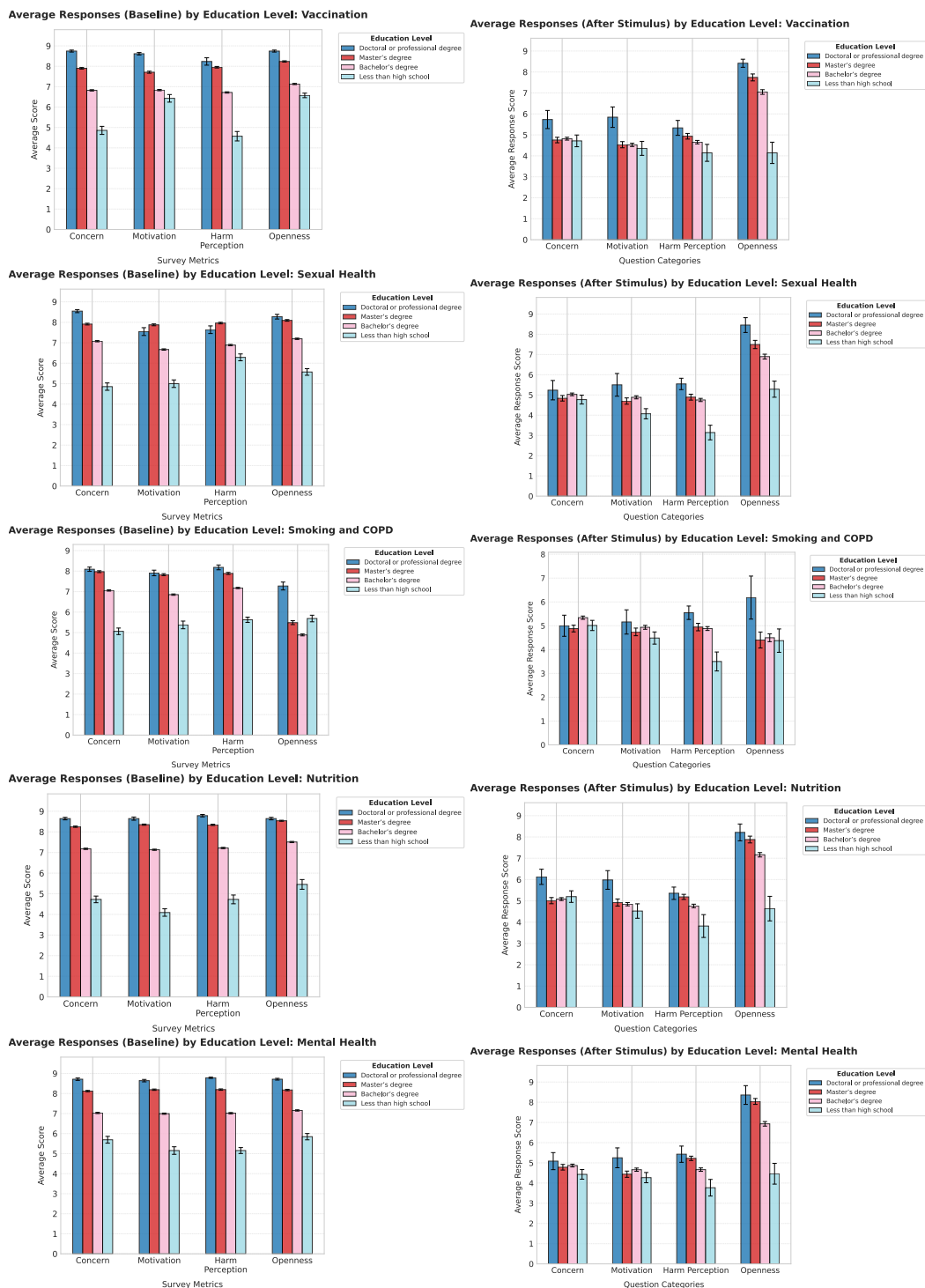


Figure 12: **Average Responses By Education Levels:** Concern, Motivation, Harm Perception and Openness of individuals with different education levels prior and after viewing any marketing content. People with a higher degree such as masters and doctoral tend to be more concerned about health concerns, and more motivated or open to practice healthy behaviors. Interestingly, individuals with less than high school are less open to smoking than other educational groups, but less concerned about the health risks related to it and less motivated to not smoke or abuse substances.



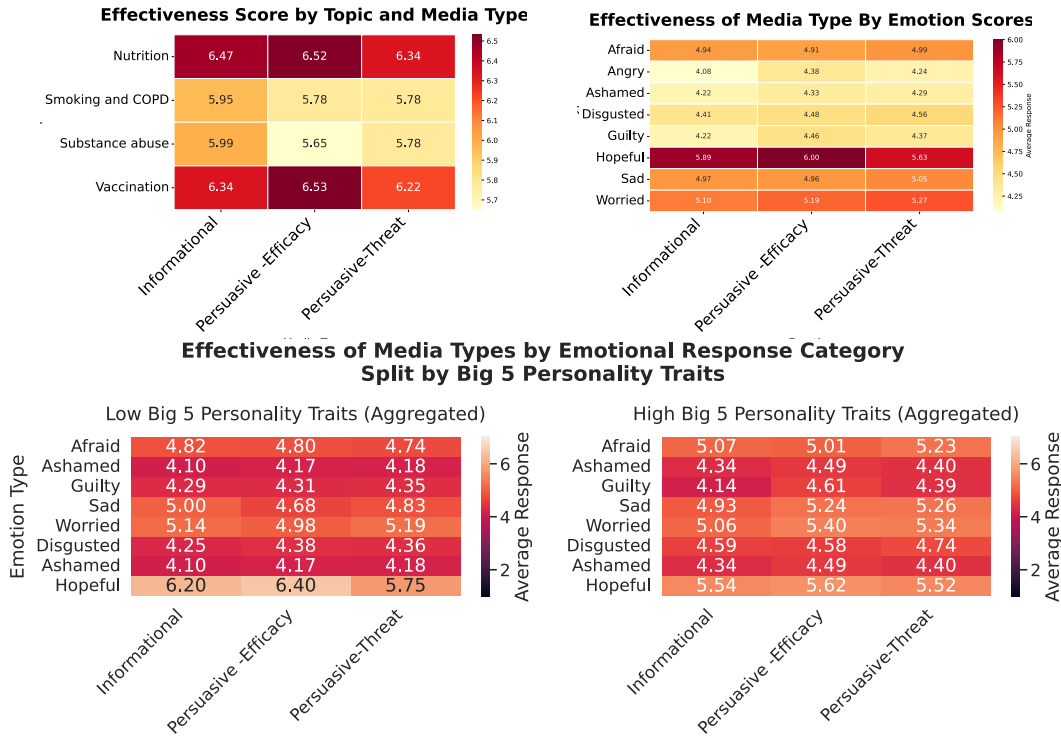


Figure 13: **Motivation and Needs for PHORECAST:** We analyze the concern, motivation, openness and harm perception scores induced by each campaign type to analyze their effect across public health topics. We observe that for some, like Smoking/COPD and Substance Abuse, Informational (*I*) campaigns are more effective, whereas for well-known public health concerns like nutrition, and vaccination, Persuasive-Efficacy (*PE*) campaigns tend to be more effective. Studies like [McCloughlin et al. \(2023\)](#) discuss similar results using COVID campaigns. Next, we illustrate the effectiveness of different campaign types (Persuasive-Efficacy (*PE*), Persuasive-Threat (*PT*), and Informational (*I*) for eliciting different emotional responses. Overall, *PT* messaging induces the highest levels of fear, anger, disgust, sadness, and worry, while *PE* is particularly useful for evoking hope. Further, our dataset reveals how individuals with varying personality traits respond differently to specific media approaches. We find that *PE* messaging tends to be more effective for individuals with lower Big 5 personality trait scores (e.g., agreeableness, extraversion), whereas *I* and *PT* based messaging are more effective for those with higher scores.

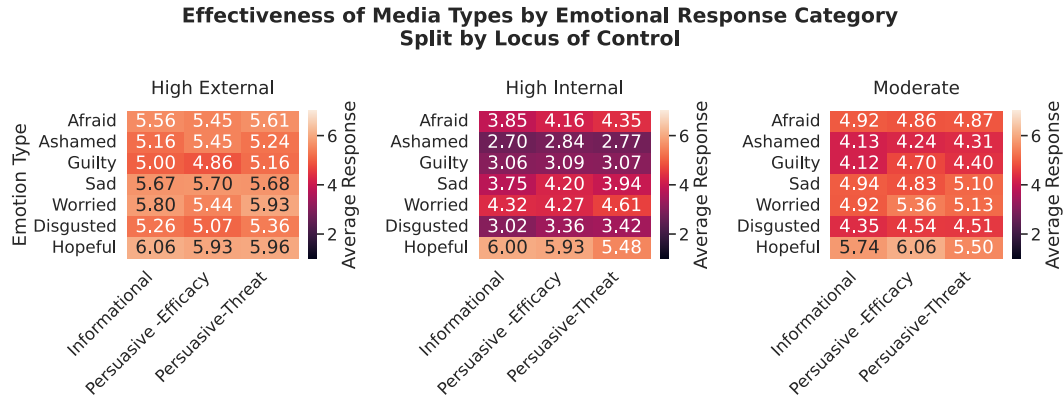


Figure 14: **Media Effect by Locus of Control:** We illustrate the effect the media type (Persuasive-Efficacy, Persuasive-Threat, and Informational) has on eliciting different emotional responses across different categories of locus of control. We observe that people with a high internal (middle) tend to feel less ashamed and guilty, while those with a high external feel the highest levels of fear, shame, guilt, sadness, worry, disgust and hope. Further, it is not clear which media type is most effective for individuals with different locus of control categories, implying that the impact of media on individuals might be moderated by their pre-existing beliefs about control over their lives.

```
User: You are a helpful assistant trained to interpret user thoughts and
feelings and predict how they would react and answer different questions
about various health topics.

Five health topics are randomly selected for you from the following list:
Nutrition, Vaccination, Mental Health, Substance abuse, COPD, Chronic
Diseases, HIV/aids, Sexual Health.

You are of the following demographics: get_demographics(row).
You have the following personality traits: get_personality(row).
You have a row['locus'].
You first answer baseline questions about each health topic.
For the topic of row['topic'], you answer as follows: get_baseline(row).
You are then shown the following image and you answer the following: [Q/As].

Given the question: 'type in every thought that came to mind viewing this
material.' What would your response be?

Assistant: "It makes me think about..."
```

Figure 15: **Training Template Structure:** Chat format used for training, showing dynamic insertion of (1) user profiles, (2) randomly sampled health topics, and (3) task-specific response targets.

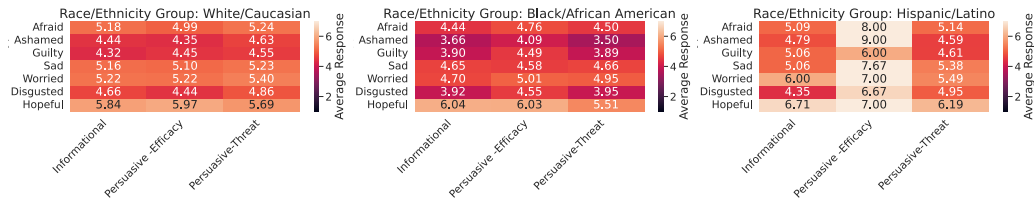
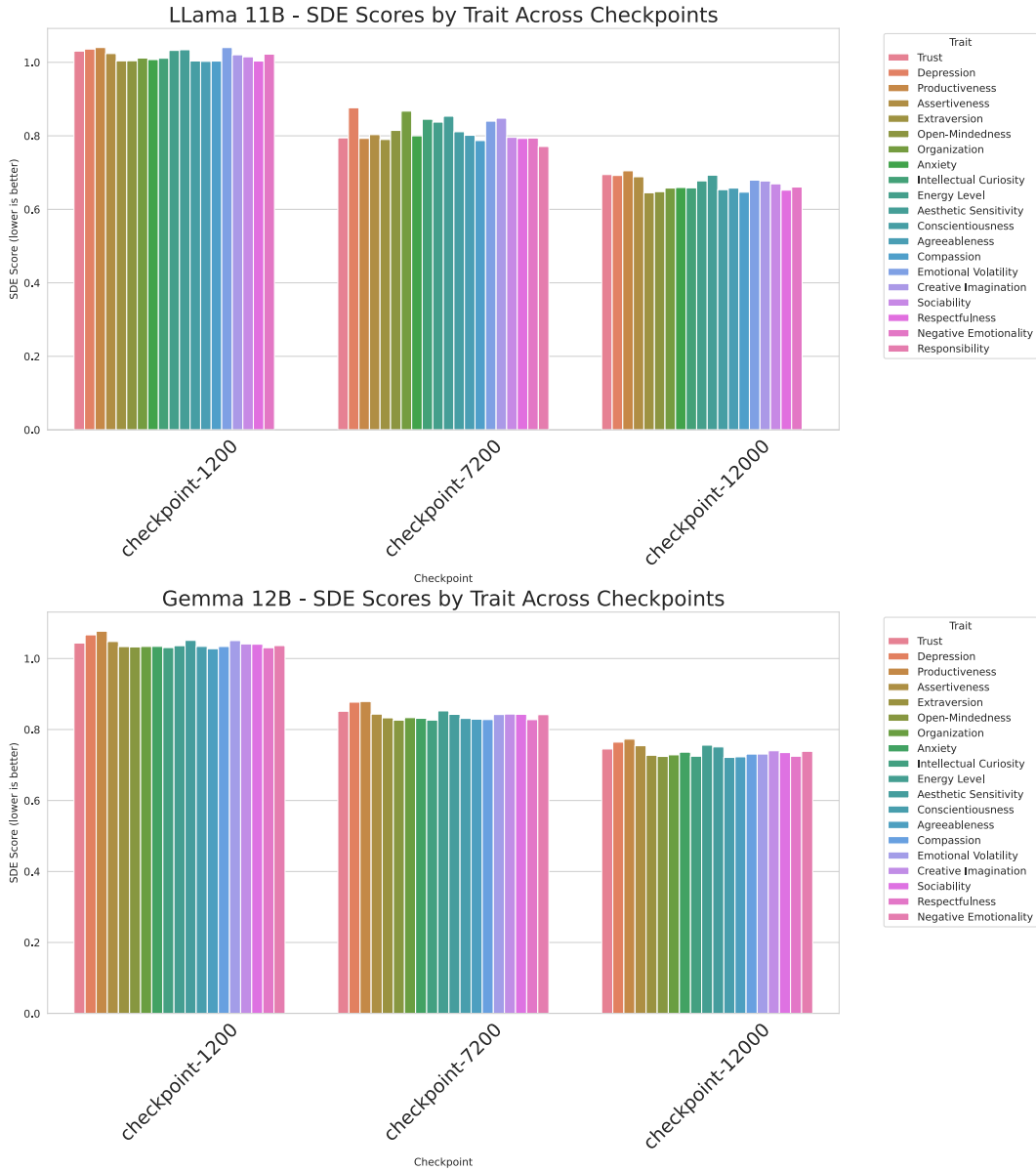


Figure 16: **Media Effect by Race/Ethnicity:** The effect the media type (Persuasive-Efficacy, Persuasive-Threat, and Informational) has on eliciting different emotional responses across different race/ethnicity groups. We observe that individuals who fall under the Hispanic/Latino group tend to react more strongly to persuasive efficacy messaging. Participants who identify as white/caucasian are likely more effected by persuasive-threat campaigns, while participants that identify as Black/African American react more strongly with persuasive campaigns.



**Figure 17: SDE Scores Across Checkpoints** of Llama (Top) and Gemma (Bottom). We observe that both models begin with a high SDE score. As the models iterate over our dataset, SDE scores continue to decrease, showing better alignment to individuals. This results in a model that not only aligns better on average but also captures the unique manifestations of different personalities with greater fidelity.

⊙ = {Christian, Man, **Black/African American**, Conservative, Married, Profession: **Skilled Trades/Labor**, Income: **\$25,000-\$49,999**, Low Internal Locus of Control/High External..}

True Response:

Health risks: I realized that long-term consumption of junk food could lead to serious health problems such as heart disease and diabetes

Gemma Before Training:

Man, that's a heavy image. It's like, McDonald's is part of the problem, right? Like, it's messing with people's health. It's kinda sad, seeing the heart monitor and the McDonald's logo together. It's like, you gotta be careful what you eat. It's a reminder to take care of yourself. It's a good message, I guess. It's like, don't let fast food control your life. It's a good thing to be healthy.

Gemma Trained:

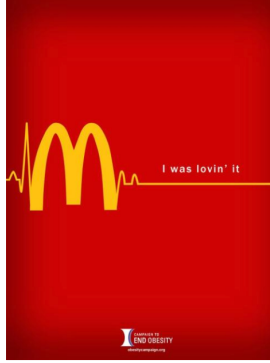
End Obesity

Llama Before Training:

I'm not sure what this is about, but it looks like some kind of advertisement or campaign. I don't really get it, though. Is it trying to say that McDonald's is healthy or something? I'm not sure. It's just a bunch of words and a logo. I don't really understand what it's trying to say.

Llama Trained:

I was lovin' it



⊙ = {Christian, Man, **White/Caucasian**, Conservative, Married, Profession: **Business/Finance**, Income: **\$50-74,999k**, Low Internal Locus of Control/High External..}

True Response: It encourages me to track my meals and recognize patterns in my eating that might not be serving my health goals.

Gemma Before Training:

It's a clever ad. The McDonald's logo shaped like a heartbeat is a bit shocking. It's trying to say fast food is bad for your health. I wonder if it's true. It's a bit extreme. I like the way they combined the two images. It's a good way to get your attention. It's a bit sad to think about. I wonder if it's true. It's a bit shocking. It's a good way to get your attention.

Gemma Trained:

I was surprised when I realized how quickly I went from being slim to obese because of my love for junk food. This change made me realize that my eating habits have a huge impact on my weight

Llama Before Training:

I'm not sure what to make of this. It seems like a weird advertisement for McDonald's, but the slogan "I was lovin' it" is catchy. I'm not sure if it's trying to be funny or serious. The image of the golden arches and the heartbeat line is kind of confusing. I'm not sure what it's supposed to represent. Maybe it's trying to say that eating at McDonald's is a good thing? But I'm not sure. I'm a bit skeptical.

Llama Trained:

I feel inspired to share this with others.

⊙ = {Bhuddist, Female, **White/Caucasian**, Conservative, Married, Profession: **Education**, Income: **\$50,000-\$74,999**, Moderate Internal / External Locus of Control..}

True Response: I wish more people understood mental health

Gemma Before Training:

It's a lot of imagery. The mental health one is good. The heart ones are nice. It's a positive message. The hand one is a bit confusing. It's a lot to take in..

Gemma Trained:

I find the message compelling and am motivated to prioritize my mental well-being

Llama Before Training:

I'm not sure what to think about this material. It seems like it's trying to convey some important messages, but I'm not really sure what they are. I'm a little overwhelmed by all the different images and words. I think I need to take a step back and try to process everything before I can form any thoughts

Llama Trained:

The importance of open conversations about mental health is highlighted throughout



⊙ = {Christian, Man, **White/Caucasian**, Conservative, Married, Profession: **Technology/Engineering**, Income: **\$75,000-\$99,999**, Low Internal Locus of Control/High External..}

True Response: Having a great mental health is very significant

Gemma Before Training:

Mental health is important. It's good to see resources available. It's a complex issue. It's good to be supportive. It's good to be understanding. It's good to be compassionate. It's good to be open-minded. It's good to be respectful. It's good to be responsible. It's good to be healthy.

Gemma Trained:

The message makes me to determine that I will prioritize my mental health and seek support

Llama Before Training:

I'm not sure what to think about this material. It seems like it's trying to convey some kind of message or idea, but I'm not really sure what it is. The images are colorful and visually appealing, but they don't seem to be saying anything specific. I'm a bit confused.

Llama Trained:

Mental health is important. I agree with this.

Figure 18: **Qualitative Examples:** We compare responses of different individuals using one of the public health campaigns related to nutrition (top), and mental health (bottom), before and after training with our dataset. Before training, Gemma is able to attempt the task of emulating the different individuals, while Llama just tries to understand and explain the visual. After training using PHORECAST, both model responses are more aligned and human-like.

## E GENERALIZING RESPONSE PREDICTIONS TO UNSEEN COMMUNICATION STRATEGIES

### E.1 PROBLEM DESCRIPTION

In practice, public health campaigns may have communication strategies that were not covered in the model’s training data. We further extend our experiments to test whether a VLM trained on a set of communication strategies can effectively generalize to health messages with novel communication strategies. The core difference from the prior use case is the *train/test partition approach*: instead of splitting randomly based on specific posters, we split based on *communication strategies*. The motivation is to evaluate whether the VLMs trained with our dataset can generalize to a new campaign message with a communication strategy not covered in the training set.

Specifically, for each VLM architecture, we trained 3 different set of weights:

1. **Set 1:** Test set includes only messages using *Self-Efficacy* strategies.
2. **Set 2:** Test set includes only messages using *Informational/Educational/Neutral* strategies.
3. **Set 3:** Test set includes only messages using *Threatening/Fear-driven* strategies.

This training and evaluation setup enables us to investigate the model’s capacity to generalize to campaign messages with unseen communication strategies in practice.

Some representative health messages for each of the communication strategies are shown in Fig. 19



(a) (b) (c) Informational/Educational/Neutral  
Self-Efficacy  
Threatening/Fear-Driven

Figure 19: **Examples of Health Campaign Message Communication Strategies.** (a) *Self-Efficacy* (Health Topic: Sexual Practice); (b) *Threatening/Fear-Driven* (Health Topic: Nutrition); (c) *Informational/Educational/Neutral* (Health Topic: Vaccination).

### E.2 EXPERIMENTS & ANALYSIS

To demonstrate the value of our dataset in improving personality- and demographic-conditioned response prediction, we compare the performance of zero-shot vision-language models (VLMs) with VLMs fine-tuned on our dataset. Both models are evaluated using the same set of system and instruction prompts to ensure a fair comparison.

#### E.2.1 PERSONALITY-SPECIFIC EVALUATIONS: GENERALIZATION TO UNSEEN “INFORMATIONAL/NEUTRAL” STRATEGY.

In this section, we evaluate whether fine-tuned VLMs can generalize to predict responses to **Informational/Neutral** health campaign messages held out during training. We compare the predicted response distributions of fine-tuned models against both zero-shot, untrained baselines and ground-truth responses.

To assess how well the models capture group-specific patterns, we investigate the true and predicted distributions across personality traits partitioned into *low*, *moderate*, and *high* levels. To ensure the statistical significance of each group-specific distribution, we excluded any personality group having fewer than 20 individuals.

Comparisons between the predicted response distribution of the zero-shot and trained Gemma model are shown in Fig. 20 for varying “Intellectual Curiosity” personality groups. After trait-conditioned training, the VLM shows substantially improved alignment with ground-truth distributions, more accurately capturing personality- and demographic-specific sentimental response patterns. In particular,  $\pm 1$  accuracy increased from 0.50 to 0.62 for the *moderate* “Intellectual Curiosity” group (a 24% gain), and from 0.45 to 0.73 for the *high* group (a 62.2% gain). Zero-shot baselines tend to give more moderate responses (6–7 out of 9), failing to capture trait-conditioned variations. In contrast, PHORECAST-trained models can capture such response distribution, such as the high-scoring (8–9) responses to Informational/Educational/Neutral messages by those *high* in Intellectual Curiosity.

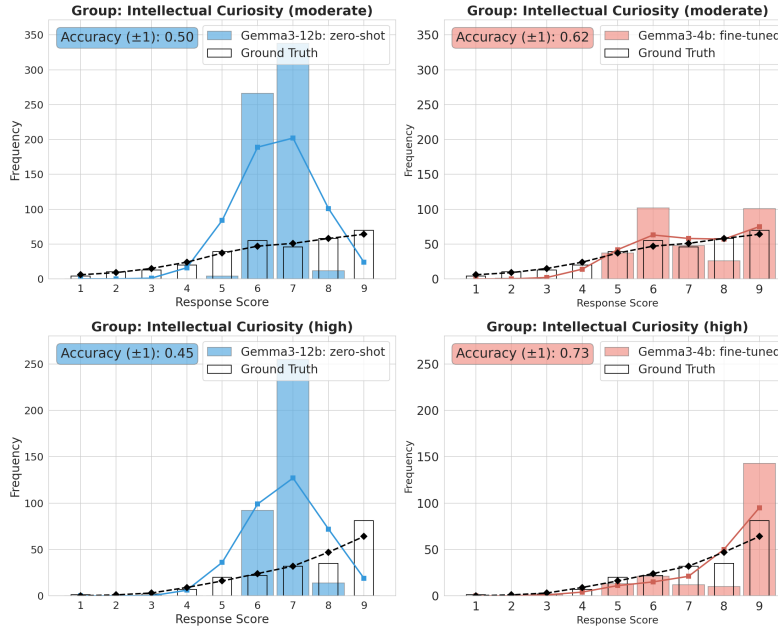


Figure 20: Comparison of sentimental response distributions from **Gemma** models on unseen *Informational/Neutral* messages, evaluated across *Intellectual Curiosity* personality groups (*moderate*, *high*). The personality group “Intellectual Curiosity: *low*” is not included since it has fewer than 20 samples in the test set. While the zero-shot model (left) shows limited sensitivity to group differences and fails to capture the true response distribution, the trained model (right) using PHORECAST closely aligns with the true personality-conditioned response patterns. Specifically, the  $\pm 1$  accuracy improved from **0.50** to **0.62** for the *moderate* “Intellectual Curiosity” group (**improved by 24%**), and from **0.45** to **0.73** for the *high* group (**improved by 62.2%**).



### E.2.2 PERSONALITY-SPECIFIC EVALUATIONS: GENERALIZATION TO UNSEEN “SELF-EFFICACY” STRATEGY.

In this section, we evaluate whether fine-tuned VLMs can generalize to predict responses to **Self-Efficacy** health campaign messages held out during training. We compare the predicted response distributions of fine-tuned models against both zero-shot, untrained baselines and ground-truth responses.

To assess how well the models capture group-specific patterns, we investigate the true and predicted distributions across personality traits partitioned into *low*, *moderate*, and *high* levels. To ensure the statistical significance of each group-specific distribution, we excluded any personality group having fewer than 20 individuals.

Comparisons between the predicted response distribution of the zero-shot and trained Gemma3 model are shown in Fig. 21 for varying “Trust” personality groups. After trait-conditioned training, the VLM shows substantially improved alignment with ground-truth distributions, more accurately capturing personality- and demographic-specific sentimental response patterns. In particular,  $\pm 1$  accuracy increased from 0.47 to 0.66 for the *moderate* “Trust” group (a 40.4% gain), and from 0.44 to 0.68 for the *high* group (a 54.5% gain).

Qualitatively, the zero-shot, pretrained baselines tend to give more moderate responses (6–7 out of 9) with some highly positive responses (8 out of 9). It fails to correctly capture the tendency to have very positive sentimental responses (9 out of 9) among the *high* “Trust” group. In contrast, PHORECAST-trained models can capture such response distribution and patterns for both *moderate* and *high* “Trust” groups. Since the sentimental response score are often imprecise by nature, there is practically little difference between prediction sentimental score of 8 versus 9 (or 5 versus 6), showing that the trained VLM is able to capture the overall trend in different “Trust” groups, as reflected in the similar distribution shapes and high  $\pm 1$  accuracy.

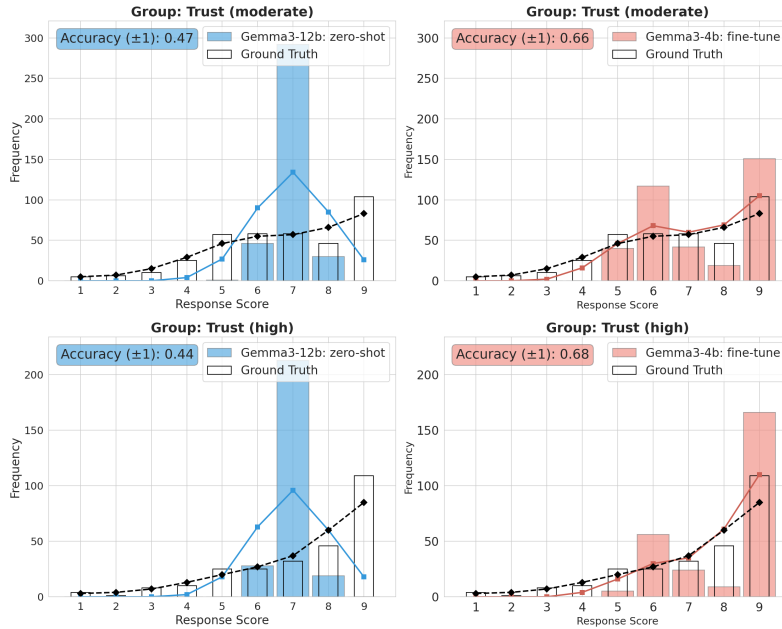


Figure 21: Comparison of sentimental response distributions from **Gemma3** models on unseen *Self-Efficacy* messages, evaluated across *Trust* personality groups (*moderate*, *high*). The personality group “Trust: *low*” is not included since it has fewer than 20 samples in the test set. While the zero-shot model (left) shows limited sensitivity to the group differences and fails to capture the true response distribution, the trained model using PHORECAST (right) closely aligns with the true personality-conditioned response patterns. Specifically, the  $\pm 1$  accuracy improved from **0.47** to **0.66** for the *moderate* “Trust” group (**improved by 40.4%**), and from **0.44** to **0.68** for the *high* group (**improved by 54.5%**).

### E.2.3 PERSONALITY-SPECIFIC EVALUATIONS: GENERALIZATION TO UNSEEN “THREATENING/FEAR-DRIVEN” STRATEGY.

In this section, we evaluate whether fine-tuned VLMs can generalize to predict responses to **Threatening/Fear-driven** health campaign messages held out during training. We compare the predicted response distributions of fine-tuned models against both zero-shot, untrained baselines and ground-truth responses.

To assess how well the models capture group-specific patterns, we investigate the true and predicted distributions across personality traits partitioned into *low*, *moderate*, and *high* levels. To ensure the statistical significance of each group-specific distribution, we excluded any personality group having fewer than 20 individuals.

Comparisons between the predicted response distribution of the zero-shot and trained Gemma3 model are shown in Fig. 22 for varying “Neurocitism” personality groups. After trait-conditioned training, the VLM shows substantially improved alignment with ground-truth distributions, more accurately capturing personality- and demographic-specific sentimental response patterns. In particular,  $\pm 1$  accuracy increased from 0.51 to 0.67 for the *moderate* “Neurocitism” group (a 31.4% gain), and from 0.51 to 0.70 for the *high* group (a 37.3% gain).

The zero-shot pretrained baseline again shows poor alignment with true sentimental responses, predominantly predicting moderate scores of 6–7 out of 9. Additionally, its predicted response distribution shows little variation between the *moderate* and *high* neuroticism groups. In contrast, the PHORECAST-trained model captures sentiment variations both within and between different “Neurocitism” personality groups.

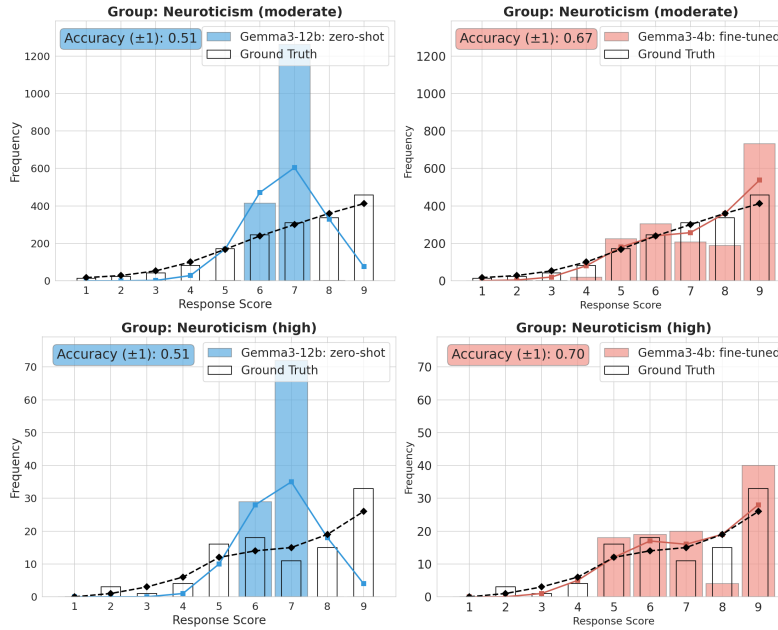


Figure 22: Comparison of sentimental response distributions from **Gemma3** models on unseen *Threatening/Fear-driven* messages, evaluated across *Neurocitism* personality groups (*moderate*, *high*). While the zero-shot model (left) shows limited sensitivity to the group differences and fails to capture the true response distribution, the trained model using PHORECAST (right) closely aligns with the true personality-conditioned response patterns. Specifically, the  $\pm 1$  accuracy improved from **0.51** to **0.67** for the *moderate* “Neurocitism” group (**improved by 31.4%**), and from **0.51** to **0.70** for the *high* group (**improved by 37.3%**).

## F FUTURE PRACTICAL USE CASE: VLM-ENABLED HEALTH COMMUNICATION STRATEGY RECOMMENDATION

### F.1 PROBLEM DESCRIPTION AND PREDICTION PIPELINE

In this section, we describe a potential practical use case of VLM-enabled trait-conditioned response prediction as an interesting line of future work: VLM-enabled communication strategies recommendation tailored to specific personality or demographic groups. Based on predicted responses to different health campaign posters, we can aggregate VLM-predicted reactions across individuals within a group to identify the most effective messaging strategy. An overview of the VLM-enabled recommendation pipeline is shown in Fig. 23.

The high-level idea is to aggregate the responses of each person in a group  $G$  to different communication strategies in the targeted health topic. Given a health topic for which behavior change is targeted, we consider a set of health campaign messages  $V^s$ , each associated with a communication strategy  $s \in \text{Threatening/Fear-driven, Self-Efficacy, Informational/Neutral}$ . Using a trained VLM, we predict how individuals with given traits respond to each strategy:

$$\hat{y}^{(i),s} = \text{VLM}(V^s, x_{\text{persona}}^{(i)}, x_{\text{demo}}^{(i)}) \quad \forall i \in [1, N]$$

where  $\hat{y}^{(i)}$  is the individual  $i$ 's behavioral response to visual health campaign message  $V^s$ , conditioned on their personality  $x_{\text{persona}}^{(i)}$  and demographic information  $x_{\text{demo}}^{(i)}$ , and  $N$  is the total individuals in group or community  $G$ . The behavioral responses of the group  $G$  to different health messages  $V^s$  with different strategies  $s$  are aggregated as:

$$y^{\text{group},s} = \sum \hat{y}^{(i),s} / N$$

As discussed in the previous sections, the behavioral responses in PHORECAST are measured on a 9-point Likert Scale, in which scores  $\geq 7$  correspond to "positive" responses. Therefore, to recommend the best communication strategies for a group/community  $G$ , the VLM outputs all communication strategies  $s$  such that  $y^{\text{group},s} \geq 7$ .

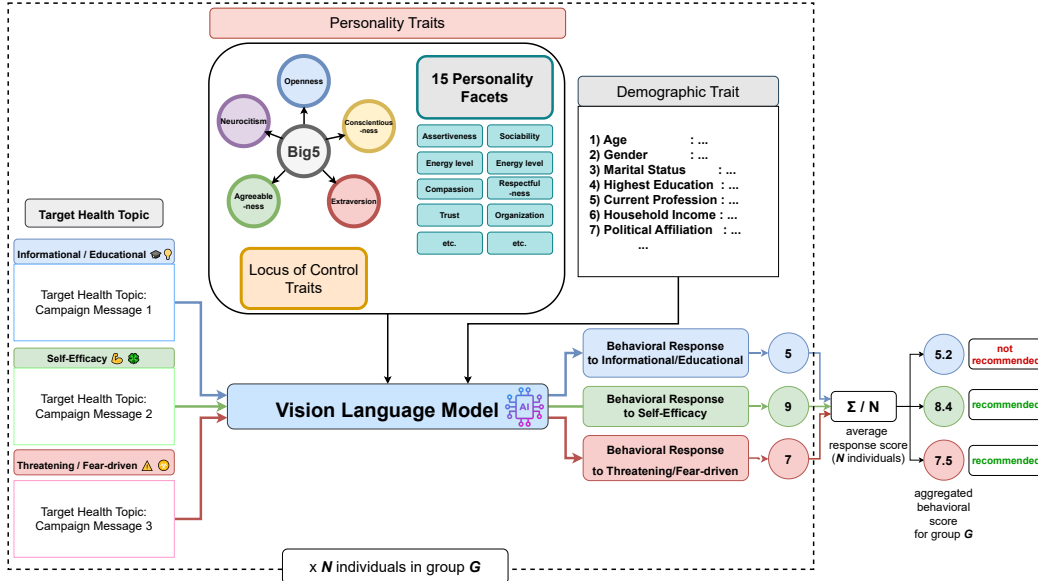


Figure 23: VLM-enabled Personality- and Demographic-conditioned Communication Strategy Recommendation to different Individuals/Communities. The model aims to recommend the potentially most effective communication strategy that likely have positive impacts on a given individual/community with a particular personality and demographic traits in a health topic.

By leveraging a VLM-enabled communication strategy recommendation system, we can tailor public health messages to specific personality and demographic profiles, thereby maximizing message

1944 effectiveness for diverse target groups. This line of future work has the potential to enhance awareness  
1945 of health issues, spread important health information, and promote healthier behaviors at scale.  
1946 Beyond the immediate applications in public health, this framework is also applicable to future  
1947 research in other disciplines such as political science, education, and social marketing. These  
1948 applications not only highlight interesting lines of future technical works but also the direct societal  
1949 impacts of our PHORECAST dataset and the models presented in this paper.

1950 To account for multiple effective strategies, we suggest using an evaluation pipeline that considers  
1951 any recommended strategies to be "correct" if it is among the true effective strategies. For example, if  
1952 all strategies for the health topic "Nutrition" are effective for personality "Open-Mindedness: *high*",  
1953 any recommended strategy in "Nutrition" is considered to be valid for this personality group.  
1954

1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997