CLAAS: Conditional LVLM-enabled Attribute-Aware response Simulation for Public Health Education

Anh N. Nhu[1], **Jenna Yang**[2], **Laura Zheng**[1,3*],
Sylvette La Touche-Howard[2], **Tracy M. Zeeger**[2], **Ming C. Lin**[1]
[1]Department of Computer Science, University of Maryland, College Park
[2]School of Public Health, University of Maryland, College Park
[3]Waymo
Correspondence: anhu@umd.edu

Abstract

Modeling how diverse communities respond to public health messages is critical for training students to design targeted communication materials. However, no method is currently available to iteratively evaluate how different populations may perceive their campaign designs before human-subject testing. Despite progress in VLM/LLM-based psychometric simulation, community simulators for public health education remain largely unexplored. We reframe VLMs as attribute-aware community response simulators and introduce CLAAS, to our knowledge the first multimodal response simulation tool for public-health higher education. CLAAS aligns open-source VLMs, such as `Gemma3`, `Llama-3.2`, and `Qwen2.5-VL`, with an in-house survey dataset of 33K+ responses from 1K+ participants to simulate trait-conditioned emotional, behavioral, and sentimental responses. Quantitatively, attribute-aligned VLMs consistently outperform zero-shot and random baselines and generalize to unseen communication strategies. Beyond dataset evaluation, our pilot studies in real course sessions across two semesters demonstrate CLAAS's potential practical relevance in educational settings. Most students reported CLAAS provided useful feedback and generated responses aligned with their expectations. This study suggests attribute-aware VLM support experiential public health education and promote a socially impactful research direction: interactive population simulator for community-centered communication design.

1 Introduction

Public health education helps students understand how diverse communities perceive, interpret, and respond to health campaign messages to design health posters that effectively promote awareness, shape individuals' opinions, and motivate healthier behaviors at scale (Adegoke et al., 2024). Although students are trained to account for demographic diversity, behavioral factors, and social context when designing health messages, their exposure is often limited to static case studies or hypothetical scenarios. As a result, public health students have limited opportunities to experiment with how different populations may respond to different health messages before entering practice, especially if there are risks of adverse effects.[1] [2]. Individual attitudes around health issues, such as vaccination, diet, and vascular health, are influenced by multiple interweaving factors. Societal views, individual personality, lived experience, and perceptions of behavioral control create a wide spectrum of responses to health communication messages (Taboada-Vázquez et al., 2021; Olsacher et al., 2023; Nah et al., 2023). Evaluating health communication strategies in real communities is costly and time-intensive, thereby motivating the need for efficient and reliable response simulation methods to support experiential learning and assessment of health messages. A recurring challenge for public health education is the lack of practical classroom tools to iteratively test and evaluate how different communities, represented by the respective demographic compositions of individuals and each of them further characterized by psychometric and behavioral traits, respond to various visual messages for public health campaigns.

In this paper, we introduce a novel human-centered NLP problem formulation: attribute-aligned community response simulation for public health communication education. Rather than proposing a new architecture or optimization method, our contribution is to reframe VLMs as attribute-aware population simulators that model how diverse communities may respond to visual health messages. Specifically, our core contributions are: **(1)** we formulate community response simulation as a *trait-conditioned multimodal prediction task over sentimental, behavioral, and emo-*

[1]* Work done while being at the University of Maryland.
[2]Code and website at `github.com/anh-nn01/claas`.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 45706–45746
October 24-29, 2026 ©2026 Association for Computational Linguistics

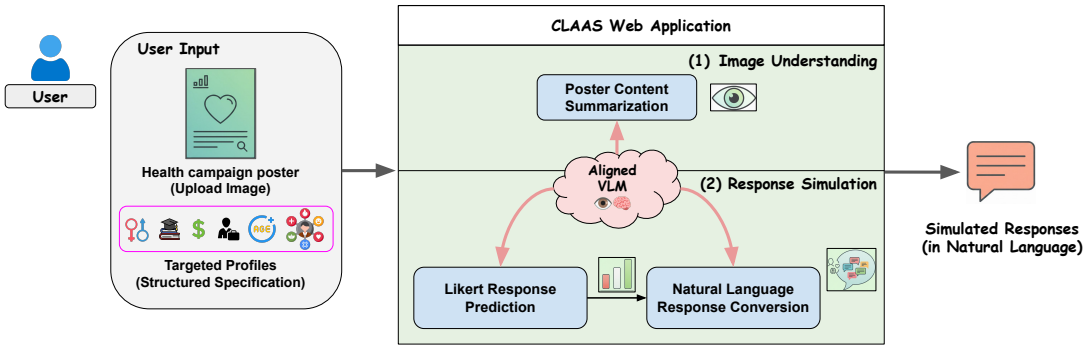


Figure 1: CLAAS: Conditional LVLM-enabled Attribute-Aware response Simulation

tional Likert responses, capturing both *individual variation* and *population-level response* patterns; **(2)** we introduce an attribute-aware alignment framework that adapts open-source VLMs to simulate responses conditioned on *demographic* and *psychometric* traits; and **(3)** we evaluate this framework in a practical public health education setting, including cross-communication-strategy generalization and real-classroom pilots.

We implemented this attribute-aligned VLMs within a novel educational software system that simulates community responses with individualized feedback to student-designed public health posters in natural language. Our in-classroom pilot tests of CLAAS (N=32) revealed that a large majority of students (**87.5**% rated ≥ 3 out of 5) reported that the system **(a)** provided useful feedback/evaluation of their campaign message designs and **(b)** generated responses perceived as realistic and aligned with their expectations. These findings suggest the potential of VLMs/LLMs as attribute-aware response simulation in modeling sentiment, behavior, and emotion for public health education and training. Together, these results suggest that trait-aligned VLM-based response simulation can facilitate more reflective and experiential learning by allowing students to assess and understand how their poster designs may impact various communities of diverse demographic factors with different individual psychometric and behavioral traits[3].

2 Related Works

2.1 Attribute-aware LLM-enabled Simulation

LLMs have demonstrated strong promise in simulating personality-driven and population-level human behavior across interactive and social settings. Park et al. (2026) introduces a large-scale generative agent simulation framework in which LLM-based agents are instantiated with memory, goals, and social context to model the behaviors and interactions of 1,000 individuals. Klinkert et al. (2024) evaluates LLMs' capability to emulate human personality traits in interactive settings, showing that LLMs can produce personality-consistent behaviors across repeated interactions. Wang et al. (2025) systematically assesses whether LLMs can emulate human personality traits using established psychological inventories. Sorokovikova et al. (2024) provides further experiments of Big Five personality simulation in LLMs. Jiang et al. (2024) shows that LLM outputs can be influenced by explicit LLMs to simulate human decision-making and behavior, demonstrating their potential as behavioral simulators. Serapio-García et al. (2025) shows that the base open-source LLMs exhibit stable intrinsic personalities and are often resistant to personality-conditioned prompting. In a related application, Magnossão de Paula et al. (2025) incorporates demographic factors into LLM for a more complete sexism detection from different perspectives. To further incorporate psychological behavior patterns into language models at the foundation level, Vu et al. (2025) propose PsychAdapter as a method enabling the development of chatbots with specific personality profiles and psychological traits. Via the analysis of social media texts (Bui et al., 2025; Espinosa and Salathé, 2024; Kwok et al., 2024), LLM/VLM is used to predict infectious disease outbreaks, facilitate health screenings, and understand public sentiment on current health issues for messaging in English and low-resource languages.

2.2 VLM-enabled Simulation for Education

There is a growing interest in properly integrating LLMs into health-related education, with recent work examining both their benefits and limitations

[3]Throughout this paper, we use the terms "attributes" and "traits" interchangeably to refer to demographic, psychometric, and behavioral characteristics used to condition simulated diverse communities consisting of varying individual responses.

in instructional settings. Baglivo et al. (2023) suggests LLMs can support learning and reasoning in health and medical science. Prior work further highlights that effective AI integration in education requires not only technical exposure, but also environments that emphasize evaluation, reflection, and responsible use (Naamati-Schneider, 2024). These findings, along with the current lack of such tools in public health, motivate the development of VLM-based interactive educational tools explicitly designed to enhance experiential learning in public health education.

In Medicine, VLM/LLM are used to train medical students in patient conversations and to answer basic questions patients may have about a specific health condition (Weisman et al., 2025). Hu et al. (2025) proposes a population-aligned persona generation framework for LLM-based social simulation, emphasizing the importance of matching simulated trait distributions to real-world demographic and psychometric data. Rende et al. (2025) show that personality-conditioned LLM-based agents can reproduce nuanced human behaviors in a heating preference control task, suggesting that trait-conditioned language models can capture behavioral variation across domains.

However, *the potential of LLMs and VLMs* to **assess the attribute-aware impact of visual health communication messages**, at both **individual** and **population** levels **and** to **provide experiential education and training for public-health students and practioners**, *remains underexplored.*

3 Methodology

In this section, we describe the alignment data, training, and model pipelines for **(1)** VLM-enabled individual/community response prediction and **(2)** its use for interactive classroom education.



(a) Fear-Driven (b) Self-Efficacy (c) Neutral

Figure 2: **Examples of 3 different Communication Styles in Health Campaign Messaging.** (a) Threatening/Fear-Driven (left); (b) Self-Efficacy (center); (c) Informational/Educational/Neutral (right).

3.1 Alignment Data

To support evaluation and trait-aware alignment training of VLMs, we leverage large-scale in-house public health study data collected by the School of Public Health at the University of Maryland, College Park (Qadri et al., 2025). The dataset includes more than 1,000 valid participants spanning diverse demographic (e.g., education, income, occupation) and personality traits (e.g., Big Five, curiosity, assertiveness, trust, locus of control). It contains over 30,000 responses to 36 visual public health campaign posters across 9 health topics: sexual practice, substance abuse, smoking and COPD, heart disease, cystic fibrosis, vaccination, HIV, nutrition, and mental health.

Communication strategies are broadly grouped into three categories: **(a)** Threatening/Fear-driven, **(b)** Self-Efficacy, and **(c)** Informational/Educational/Neutral, as shown in Figure 2. Each participant was given 2–3 randomly selected posters and provided 12x Likert-scale (1–9) sentiment/emotion/impact ratings and one free-text response per poster. The survey questions were designed by public health and HCI experts, who are faculty members and principal investigators in their respective domains. Please see Appendices D, E, F for details on survey, data collection and quality assurance.

3.2 VLM-enabled Attribute-Aware Response Simulation

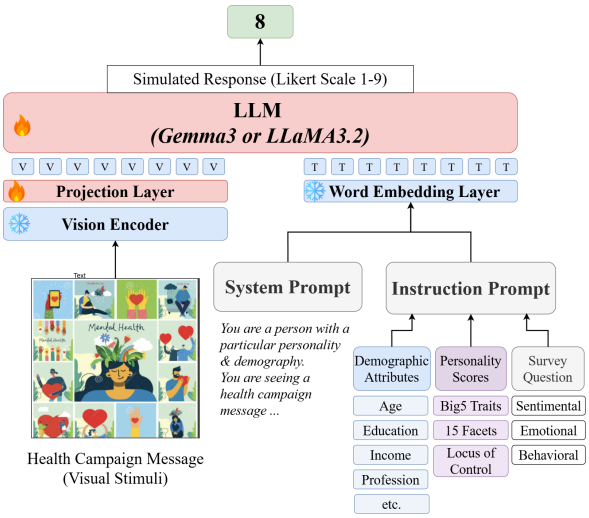


Figure 3: **Model Alignment Pipeline for Trait-Conditioned Response Simulation.** A health campaign message serves as the visual stimulus for predicting the responses. Demographic attributes and personality traits condition the simulated individual or community. Target responses are either a 9-point (1 = strongly disagree, 9 = strongly agree) or 6-point (1 = strongly disagree, 6 = strongly agree) Likert scale. During alignment fine-tuning, only the projection and language model layers are updated via LoRA.

We develop and assess open-source VLMs to simulate trait-conditioned responses to visual pub-

Test Response Alignment Results On Unseen Communication Strategies

Model / Communication Strategy	Overall	Sentiment Prediction	Behavior Prediction	Emotion Prediction
"Informational/Neutral"	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2
Uniform Random Baseline	10.5 / 30.4 / 48.1	9.8 / 24.7 / 42.6	8.4 / 26.7 / 41.8	10.7 / 31.0 / 48.9
Priors Distribution-based Baseline	11.5 / 34.4 / 60.7	14.4 / 41.2 / 64.0	13.2 / 36.8 / 61.2	10.9 / 31.7 / 58.2
Llama3.2 11B Base	12.7 / 39.9 / 59.8	11.3 / 30.0 / 48.3	11.6 / 31.9 / 54.5	12.4 / 37.5 / 58.3
Llama3.2 11B Trained	**32.3 / 56.4 / 72.9**	**37.6 / 67.7 / 85.8**	**36.6 / 63.1 / 83.7**	**28.3 / 50.3 / 66.3**
Gemma3 12B Base	13.7 / 43.8 / 68.4	16.1 / 48.2 / 84.6	13.2 / 43.8 / 74.4	11.1 / 37.6 / 61.6
Gemma3 4B Trained	**30.4 / 57.3 / 73.6**	**37.8 / 68.3 / 83.5**	**36.7 / 65.5 / 83.6**	**26.4 / 52.0 / 68.1**
Qwen2.5-VL 7B Base	16.1 / 36.7 / 53.0	12.4 / 38.6 / 57.9	11.3 / 38.9 / 61.1	18.0 / 34.9 / 48.2
Qwen2.5-VL 7B Trained	**29.2 / 54.1 / 72.2**	**35.8 / 66.4 / 86.9**	**33.2 / 62.1 / 83.5**	**26.1 / 47.5 / 65.4**
"Self-Efficacy"	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2
Uniform Random Baseline	11.3 / 31.1 / 48.5	9.1 / 27.6 / 43.4	8.7 / 28.7 / 44.4	11.0 / 31.3 / 49.2
Priors Distribution-based Baseline	12.5 / 36.5 / 64.3	13.5 / 41.2 / 62.3	12.9 / 39.7 / 65.0	11.4 / 32.6 / 61.5
Llama3.2 11B Base	12.9 / 41.1 / 60.9	12.8 / 31.2 / 48.8	9.7 / 30.4 / 55.1	12.4 / 39.7 / 59.4
Llama3.2 11B Trained	**41.2 / 63.4 / 76.4**	**33.3 / 77.8 / 100.0**	**44.4 / 72.2 / 88.9**	**41.7 / 57.6 / 67.4**
Gemma3 12B Base	13.9 / 45.9/ 71.0	16.3 / 45.9 / **87.7**	15.2 / 46.0 / 77.0	12.2 / 41.0 / 64.5
Gemma3 4B Trained	**31.3 / 58.7 / 74.5**	**39.4 / 66.9 /** 84.2	**37.1 / 64.8 / 82.0**	**27.4 / 54.0 / 69.7**
Qwen2.5-VL 7B Base	16.4 / 35.5 / 52.6	11.2 / 31.7 / 48.8	12.0 / 38.9 / 56.9	19.0 / 35.6 / 50.2
Qwen2.5-VL 7B Trained	**30.5 / 53.1 / 70.9**	**35.0 / 67.4 / 86.0**	**31.3 / 60.6 / 80.5**	**28.1 / 48.0 / 64.8**
"Threatening/Fear-driven"	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2
Uniform Random Baseline	10.8 / 31.6 / 49.7	8.8 / 28.3 / 44.9	7.5 / 27.0 / 44.6	11.2 / 31.7 / 49.4
Priors Distribution-based Baseline	11.6 / 35.0 / 60.0	14.3 / 41.8 / 67.6	12.1 / 35.8 / 58.5	11.1 / 33.0 / 56.9
Llama3.2 11B Base	13.3 / 38.4 / 58.7	10.4 / 29.5 / 49.7	11.9 / 29.5 / 52.6	12.4 / 35.7 / 57.2
Llama3.2 11B Trained	**28.3 / 50.2 / 66.2**	**33.4 / 64.7 / 82.9**	**35.2 / 61.8 / 80.4**	**24.3 / 43.4 / 58.6**
Gemma3 12B Base	14.5 / 43.0 / 66.7	16.6 / 49.0 / **86.8**	14.1 / 41.8 / 71.7	12.1 / 38.2 / 61.3
Gemma3 4B Trained	**28.6 / 52.7 / 69.3**	**35.2 / 66.8 /** 83.8	**34.7 / 61.7 / 79.4**	**24.7 / 46.7 / 63.1**
Qwen2.5-VL 7B Base	14.9 / 35.5 / 53.3	14.3 / 43.9 / 66.8	10.1 / 35.4 / 53.9	15.8 / 33.7 / 49.0
Qwen2.5-VL 7B Trained	**27.6 / 50.2 / 68.1**	**33.3 / 62.9 / 81.0**	**33.9 / 58.8 / 79.6**	**23.3 / 44.0 / 62.1**

Table 1: Model generalization results (exact / ±1 / ±2 tolerance) for each response type on *unseen* messaging strategies. The ±1 and ±2 metrics account for the imprecise nature of Likert responses. Each trained model was tested on posters whose strategy was *excluded* from its training set. Attribute-aware alignment improves overall ±1 tolerance by **18.6%–32.3%** over the random baseline and **9.7%–22.3%** over zero-shot counterparts (the strongest baselines). The largest gains occur in *sentiment prediction*, with ±1 improvements of **36.4%–50.2%** and ±2 improvements of **38.0%–56.6%**. Both one-side McNemar and Cluster-robust pair boostrap statistical tests ($B = 10{,}000$, 95% CI) show that these ***gains are statistically significant at*** $p \leq 0.0001$ across model backbones, unseen strategies, and metrics, demonstrating that ***response-aligned models generalize to novel health message designs and can simulate attribute-conditioned community responses***.

lic health posters under unseen communication strategies, which is at the core of our proposed CLAAS Interactive Response Simulation Tool. Response emulation is conditioned on targeted *population demographic* and *individual personality* attributes (see Sec. 3.1). The target output consists of Likert-scale responses to survey questions capturing different aspects of the simulated respondents' perceptions of the message in the following:

- **Sentiment** (increased concern / awareness about the health topic). Example:

```
The message makes me more concerned about the health risks of heart disease. - Scale: 1 (not at all) - 9 (extremely)
```

- **Behavior** (motivation to avoid harmful actions or adopt healthy habits). Example:

```
The message motivates me to make healthy eating choices - Scale: 1 (not at all) - 9 (extremely)
```

- **Emotion** (e.g., Angry, Afraid, Guilty, Sad, Ashamed, Disgusted, Hopeful). Example:

```
To what extent did the material make you feel: Scale: 1 (not at all) - 9 (extremely) - Worried
```

Additional examples are provided in Appendix F.

To guide simulation, a system prompt defines the VLM's role, while personality and demographic in-

formation are embedded in the instruction prompt to contextualize the modeled individual or simulated community. The alignment fine-tuning and response simulation pipeline is illustrated in Figure 3. We employ Low-Rank Adaptation (LoRA) (Hu et al., 2021) for parameter-efficient fine-tuning of each component. For zero-shot models, the architecture remains unchanged with all weights frozen. Prompt templates used for both zero-shot and fine-tuned models are provided in Appendix A.

3.3 Interactive Attribute-Aware Responses

To assess the quality and usefulness of CLAAS for experiential public health training, we developed a web-based interactive application built around attribute-aligned VLMs. The interface allows students to upload their designed health campaign posters and specify target demographic and personality profiles via structured controls, enabling interactive simulation of community responses. As described in Section 5, trait-aligned VLMs are trained and evaluated using Likert-scale responses for precise quantitative assessment. However, structured Likert-scale outputs alone are insufficient for experiential learning, motivating the need for more naturalistic responses. Therefore, we employ a two-stage conditional response generation approach that converts predicted Likert-scale responses into natural-language responses with varying tones and lexical styles, as shown in Figure 1. This process involves prompting the model to translate the simulated Likert-scale responses into conversational content that reflects the simulated responses. To further support user trust and interpretability, the model first generates a summary of its understanding of the uploaded poster before producing the simulated response. Please refer to Appendix A and our code for further details of our prompting methods.

4 Experimental Setup and In-Class Pilot

Human-Centered and Domain-Grounded Design. CLAAS was developed through a multidisciplinary collaboration with public health faculty and HCI researchers, who motivated the educational need, guided survey and attribute design, and shaped the classroom deployment protocol. These domain experts designed evaluation in authentic junior/senior-level public health courses, where students used the system to receive simulated community responses to their own campaign posters.

VLM-enabled Response Simulation. To assess models' ability to simulate attribute-grounded responses, we quantitatively evaluate VLM performance against ground-truth Likert-scale responses on unseen posters with unseen communication strategies. We have three general communication strategies to test the fine-tuned model's generalization capability. For each architecture, we trained three different settings:

1. Setting 1: SFT (Supervised Fine-Tuning) on responses to *"Threatening/Fear-driven"* and *"Self-Efficacy"* posters (total $22 + 6 = 28$ posters); evaluated on *"Informational/Educational"* posters (total 8 posters).
2. Setting 2: SFT on responses to *"Threatening/Fear-driven"* and *"Informational/Educational"* posters (total $22 + 8 = 30$ posters); evaluated on *"Self-Efficacy"* posters (total 6 posters).
3. Setting 3: SFT on responses to *"Self-Efficacy"* and *"Informational/Educational"* posters ($6 + 8 = 14$ posters); evaluated on *"Threatening/Fear-driven"* posters (total 22 posters).

We use Unsloth (Daniel Han and team, 2023) to train `Gemma3:4b` (Team et al., 2025), `Llama-3.2-Vision:11B` (Grattafiori et al., 2024), and `Qwen2.5-VL:7B` (Bai et al., 2025). Results are available in Fig. 5 for Gemma3 and in Appendix G for Llama and Qwen.

Speech-Driven, Demographic-Conditioned Animated Avatars. To enhance engagement and perceived realism, we also developed an early system prototype for response delivery by using *avatar-driven, conversational speech interfaces*, as shown in Fig. 4. Please see ***Supplementary Videos*** for demonstrations and Appendix I for more details.

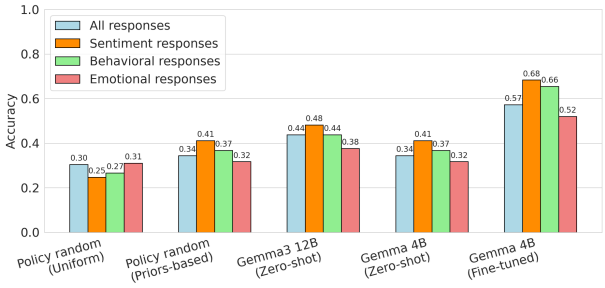


(a) Asian; Student (b) Caucasian; Healthcare

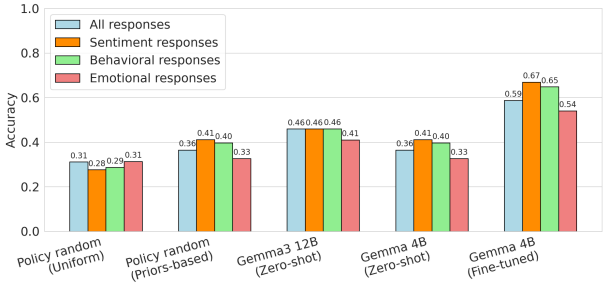
Figure 4: Animated avatars of different demographics

5 Model Evaluation & Pilot Study

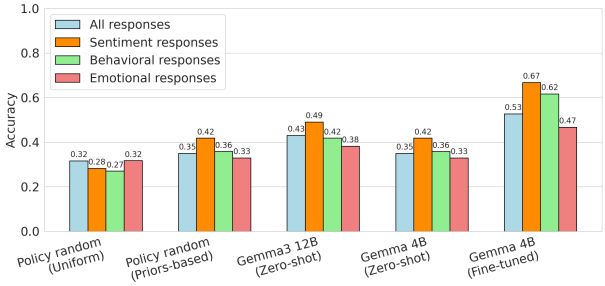
We present two sets of results: **(1)** evaluating the model's generalizability to unseen communication



(a) Informational/Educational Health Messages



(b) Self-Efficacy Health Messages



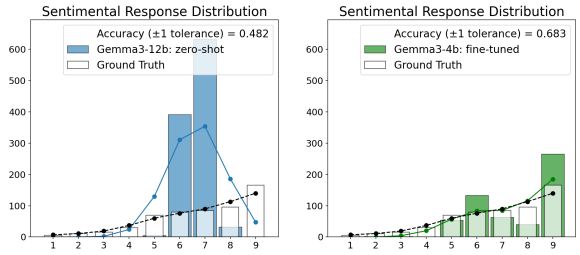
(c) Threatening/Fear-driven Health Messages

Figure 5: Gemma3 Response prediction results on unseen communication strategies. As human sentimental and emotional responses are inherently imprecise, we report accuracy using a ± 1 error tolerance w.r.t. ground-truth scores to flexibly evaluate the overall response trend. ***These gains are statistically significant at*** $p \leq 0.0001$ under exact one-side McNemar and cluster-robust pair boostrap statistical tests ($B = 10{,}000$ and 95% CI). We refer the readers to Figures 17 and 18 for additional results on Llama3.2 and Qwen2.5-VL

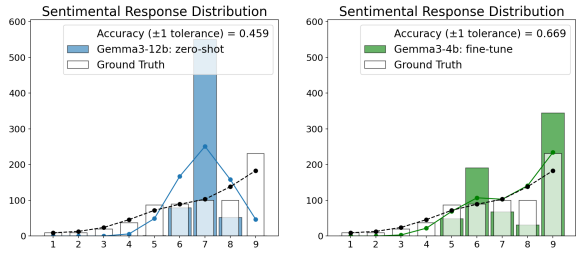
strategies in community response simulation and **(2)** measuring the effectiveness of the models in assessing the potential impact of communication strategy on novel health topics and recommending appropriate strategies for target communities.

5.1 Model Evaluation: Generalization to Unseen Communication Styles

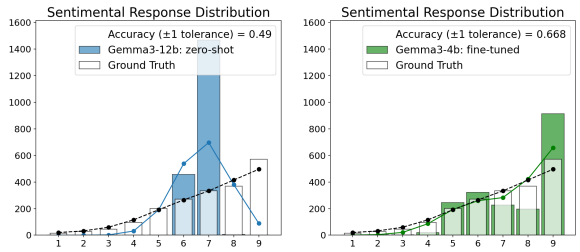
We evaluate whether aligned VLMs can generalize to predict responses to unseen communication strategies. To quantify the performance of VLMs in response simulation, we focus on two primary metrics: accuracy with a ± 1 error tolerance and accuracy with a ± 2 error tolerance. This choice is motivated by two main considerations: **(1)** sentimental and behavioral responses are inherently imprecise, and **(2)** capturing the *exact* sentiment score is not always necessary in practice. Since the primary goal is to capture the general group response distribution/trend, these relaxed accuracy metrics allow for reasonable deviation from the survey scores while still reflecting meaningful alignment. The evaluation results for `Gemma3` models are in Figure 5, with additional evaluations for both `Gemma3` and `Llama-3.2-Vision` in Table 1.



(a) Informational/Educational Health Messages



(b) Self-Efficacy Health Messages



(c) Threatening/Fear-driven Health Messages

Figure 6: Gemma3's aggregated sentimental response distributions to **unseen** communication strategies. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

Our results demonstrate that fine-tuned VLMs substantially outperform both zero-shot and random baselines across all response types – sentimental, behavioral, and emotional – in all three communication strategies. The aligned Gemma3-4b model achieves the highest performance, reaching up to 0.68 on behavioral responses to *Informational* messages and 0.67 on *Threatening/Fear-driven* messages. These gains are consistent across categories, with *improvements* of **30%-35%** in sentimental response prediction over zero-shot baselines.

We also inspect whether the improvements achieved by our attribute-aligned models over zero-

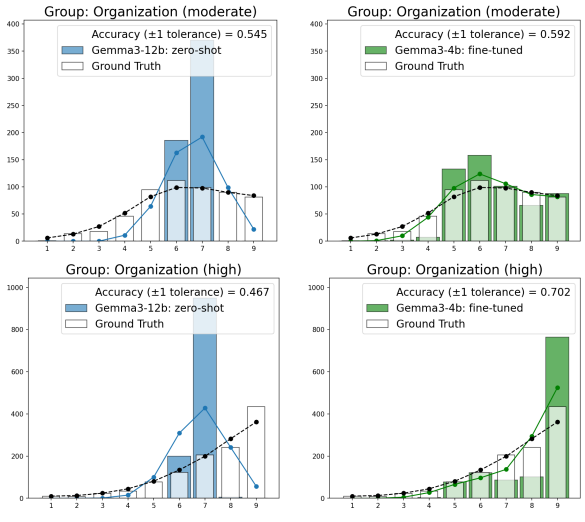


Figure 7: Sentimental response distributions of Gemma3 to unseen **Threatening/Fear-driven** posters across **"Organization" groups** (*moderate, high*). "*Low* organization" group is not included because it has < 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against ground-truth distribution.

shot baselines are statistically significant. Specifically, for each survey response i with ground truth $y_i \in \{1, \ldots, 9\}$ and model prediction $\hat{y}_i$, we define correctness under tolerance $\tau \in \{0, 1, 2\}$ as:

$$c_i = \mathbf{1}\big(|\hat{y}_i - y_i| \leq \tau\big).$$

Because survey participants answer multiple questions per poster, responses from the same person might be correlated. To account for this, we tested our models' performance gains using two complementary statistical tests:

- **Exact One-Sided McNemar Test** (McNemar, 1947): compares head-to-head wins on discordant samples where only one model is correct ($c_i^{\text{ours}} = 1, c_i^{\text{base}} = 0$ vs. $c_i^{\text{ours}} = 0, c_i^{\text{base}} = 1$).
- **Cluster-Robust Paired Bootstrap** (Efron and Tibshirani, 1994): resamples entire participant clusters with replacement ($B = 10{,}000$ iterations, 95% CI) to compute clustered p-values and confidence intervals.

We compare attribute-aligned models with zero-shot VLMs using the statistical tests above. Both statistical tests show that attribute alignment yields *statistically significant gains* ($p < 0.0001$) across all VLM backbones, held-out strategies, and metrics (including exact match accuracy), supporting attribute alignment's importance. We refer to Appendix G.1 for more statistical test details.

We investigated sentimental response distribution to unseen communication strategies. Figure 6

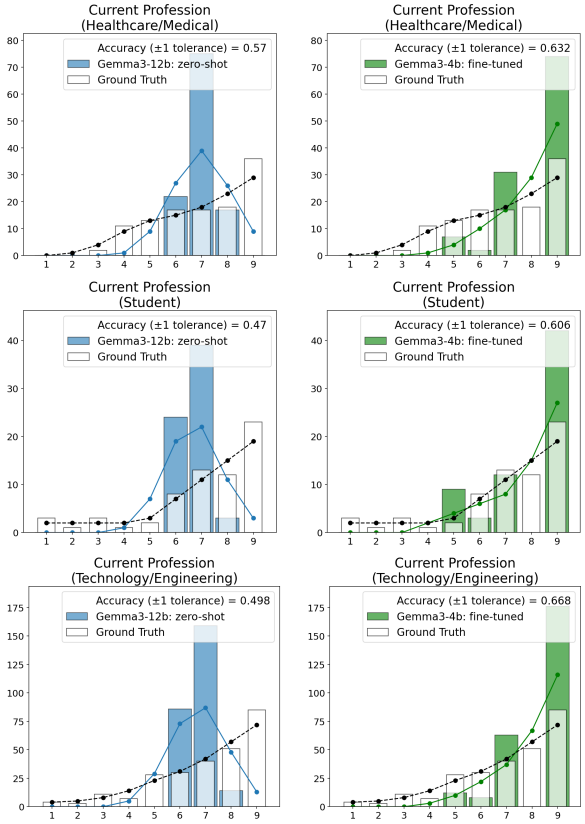
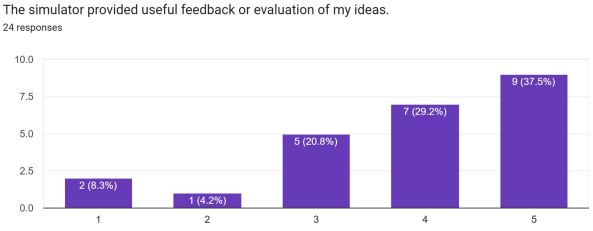
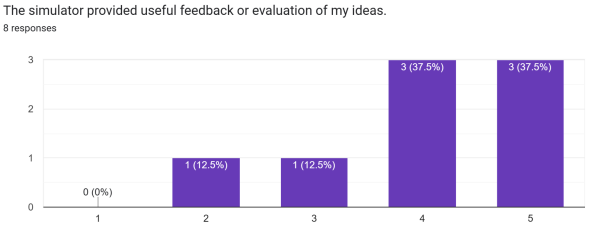


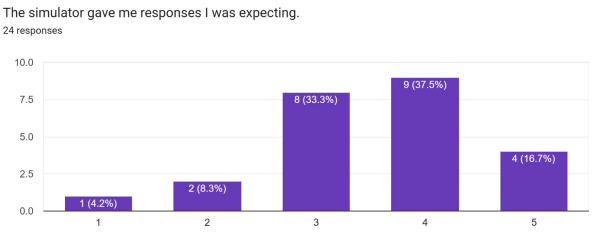
Figure 8: Sentimental response distributions of **Gemma3** models to unseen **Threatening/Fear-driven** messages across **"Current Profession" groups**. We include 3 representative profession groups and refer to Appendix G for several other groups. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

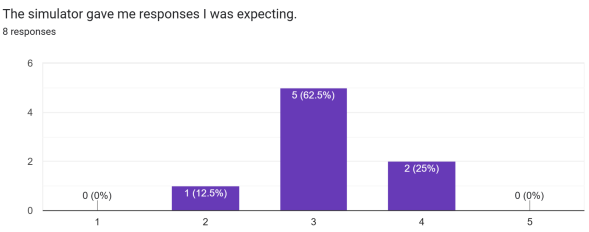
shows that aligned VLMs have significantly higher alignment with the true responses. These results underscore **the importance of domain-specific trait-conditioned alignment**: while pretrained VLMs provide a solid starting point, they fail to capture nuanced, trait-conditioned response patterns. Fine-tuned models more accurately simulate personality and demographic-conditioned human reactions, enabling more effective simulation of how different communities respond to public health campaign messages. We also conduct ablation studies in Table 3, Appendix H to evaluate how individual demographic and psychometric traits impact alignment performance. In summary, models trained with the full set of demographic and psychometric attributes generally achieve the higher accuracy. Removing either attribute group leads to noticeable performance degradation, though ablated models still outperform the zero-shot baselines. These findings indicate that demographic background and psychometric profiles contribute meaningfully to attribute-aware response prediction.



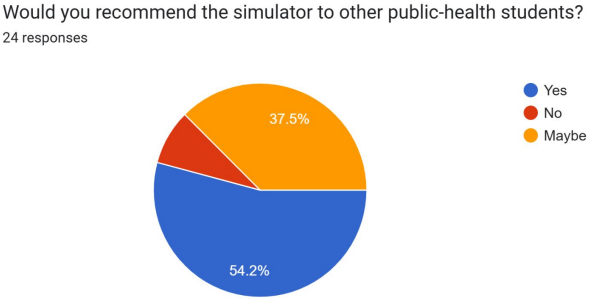


(a) Usefulness & response alignment, **Pilot Test Fall 2025**.

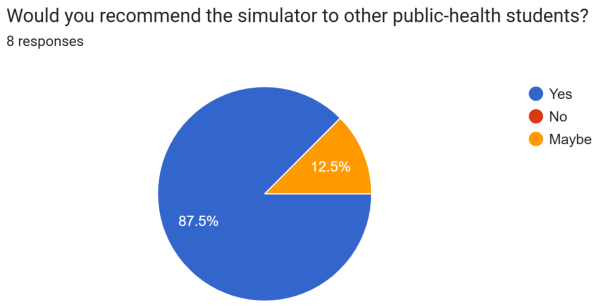




(b) Usefulness & response alignment, **Pilot Test Spring 2026**.



(c) Overall reviews, **Pilot Test Fall 2025**.



(d) Overall reviews, **Pilot Test Spring 2026**.

Figure 9: **Pilot Studies.** Students' feedback on CLAAS across Fall 2025 and Spring 2026 semesters.

5.2 Model Evaluation: Response Prediction for Targeted Groups

The zero-shot and fine-tuned Gemma3 models' simulated response distribution for different levels of "organization" traits are shown in Figure 7. For statistical significance, we only include personality groups with at least 20 individuals ($n \geq 20$), thus the *low organization* group was excluded. The zero-shot VLM over-predicts the neutral to weak positive responses (scores 6–7), failing to capture the true ground-truth distributions observed in different personality groups. In contrast, the fine-tuned model more accurately captures the true response distribution, especially in the *high-organization* group, where it captures the strong agreements (scores 8–9), demonstrating higher alignment with group-specific response patterns.

We also investigate the response distribution by demographic groups. Figure 8 also shows similar patterns for demographic-specific groups, where our model can effectively capture the overall sentiments to ***unseen*** "Threatening / Fear-driven" messages across different demographic groups. We refer to Appendix G for additional analysis of other personality and demographic groups.

5.3 In-class Pilot Study

We conducted in-class pilot studies to assess usefulness and perceived realism of CLAAS in upper-level public health communication courses across two semesters. In Fall 2025, 24 students participated; in Spring 2026, we collected 8 additional responses from the same course setting, with lower participation due to student absences. These sample sizes reflect typical enrollment and attendance constraints in junior/senior-level courses. During each session, students uploaded their own campaign posters and used CLAAS to simulate community responses under specified target profiles. This reflects intended use case of CLAAS: enabling students to compare likely community reactions to poster variants before costly human-subject studies.

Students completed a short 5-point Likert-scale survey after each session. In Fall 2025, **87.5**% gave neutral-to-positive ratings (≥ 3) on the usefulness of CLAAS for evaluating their ideas, and 87.5% rated the generated responses as aligned with their

expectations. For overall experience, 54.2% would recommend the simulator to other public health students, while another 37.5% responded "Maybe." In Spring 2026, **87.5**% of respondents indicated that they would recommend the simulator and with neutral-to-positive (≥ 3) ratings.

Despite the seemingly modest scale, our pilot studies focus on targeted domain-trained users and has provided, to our knowledge, the first real-world, human-subject evidence for an attribute-aware VLM simulator in public health communication education. The early pilot results demonstrate feasibility, educational relevance, and practical potential, motivating future larger-scale multi-year studies to evaluate learning outcomes and guide future system improvements.

6 Conclusions

In this work, we presented CLAAS the first VLM-enabled, attribute-aware response simulation system for public health communication education, which has been pilot-tested in a real-world educational setting. Our evaluation under various held-out settings show that post-training alignment of open-source VLMs generalizes to unseen campaign designs and communication strategies, capturing response variation conditioned on demographic and personality attributes. The performance gains are statistically significant at $p < 0.0001$ under both exact one-side McNemar and cluster-robust pair boostrap statistical tests (with resampling rate of $B = 10{,}000$ and 95% CI). By integrating diverse demographic factors and personality traits, the system enables trait-conditioned alignment between messaging strategies and simulated responses. Our multi-semester pilot study indicates that most ($\approx$ **87.5**%) students found the system useful for evaluating their ideas and perceived the generated responses as aligned with expectations. We plan to continue multi-year, large-scale user studies with expanded capabilities, enhance the user experience by incorporating more expressive, less monotone speech synthesis, realistic avatars with natural gestures and head movements, and more diverse, conversational responses for better user engagement.

Acknowledgement: This project has been enabled and supported by the University of Maryland Teaching and Learning Innovation Grant. We would like to acknowledge Rifaa Quadri, Swati Ramnath, Raj Bhansali, Angelo Parker, and Tom Goldstein for the web-based survey and data collection (Qadri et al., 2025) used to evaluate CLAAS framework.

Limitations

The proposed VLM-based response simulation framework leverages a large-scale web-based survey dataset with diverse demographic and psychometric profiles to demonstrate the feasibility and effectiveness of attribute-aligned response modeling to public health campaign messages and the results reflect the distribution of training datasets. The web-based studies incline to capture online users, may not fully include wide variations among all demographic groups, across rich geographic and cultural contexts. However, the proposed conceptual framework is *dataset-agnostic* and can be adapted or fine-tuned to other datasets representing different communities (of various compositions and traits) or other social applications (e.g. e-commerce, literacy, wellness, and more).

Additionally, our current evaluation focuses primarily on simulated response alignment, measured through dataset-based quantitative evaluation and students' perceived alignment with expected responses. However, the natural-language responses can be further improved for conversational delivery, including more natural phrasing, pauses, hesitations, and spoken-language style. This is especially important for **avatar-** or **conversational speech interfaces** for further improved immersion and more effective human-avatar interaction, as shown in our early system prototype examples in the Supplementary Videos and Appendix I, where response delivery may further enhance engagement and perceived realism. To improve the natural language responses, we can consider adopting close-source models, such as GPT-5.6, to convert structured Likert-scale responses into natural language using the same prompting in our current pipeline.

Ethical considerations

The system developed in this work is intended to augment experiential learning of public health students, allowing them to explore how individuals with different demographic and personality attributes may respond to health communication materials in a controlled classroom setting. We acknowledge that models capable of simulating human responses could be misused for persuasive purposes if applied outside an educational context. To mitigate this risk, CLAAS is explicitly framed as a simulation and teaching aid rather than a prescriptive or optimization tool, and system access is restricted to supervised classroom sessions only.

References

Bisola Oluwafadekemi Adegoke, Tolulope Odugbose, and Christiana Adeyemi. 2024. Harnessing big data for tailored health communication: A systematic review of impact and techniques. *International Journal of Biology and Pharmacy Research Updates*, 03(02):001–010.

Francesco Baglivo, Luigi De Angelis, Virginia Casigliani, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, and Caterina Rizzo. 2023. Exploring the possible use of ai chatbots in public health education: feasibility study. *JMIR medical education*, 9:e51421.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Nhat Bui, Giang Nguyen, Nguyen Nguyen, Bao Vo, Luan Vo, Tom Huynh, Arthur Tang, Van Nhiem Tran, Tuyen Huynh, Huy Quang Nguyen, and Minh Dinh. 2025. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, 263:108655.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Laura Espinosa and Marcel Salathé. 2024. Use of large language models as a scalable approach to understanding public health discourse. *PLOS Digital Health*, 3(10):1–15.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025. Population-aligned persona generation for llm-based social simulation. *Preprint*, arXiv:2509.10127.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Lawrence J. Klinkert, Steph Buongiorno, and Corey Clark. 2024. Evaluating the efficacy of llms to emulate realistic human personalities. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 20(1):65–75.

Kin On Kwok, Tom Huynh, Wan In Wei, Samuel Y.S. Wong, Steven Riley, and Arthur Tang. 2024. Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Computational and Structural Biotechnology Journal*, 23:3254–3257.

Angel Felipe Magnossão de Paula, J. Shane Culpepper, Alistair Moffat, Sachin Pathiyan Cherumanal, Falk Scholer, and Johanne Trippas. 2025. The effects of demographic instructions on llm personas. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3045–3049. ACM.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Lior Naamati-Schneider. 2024. Enhancing ai competence in health management: students' experiences with chatgpt as a learning tool. *BMC Medical Education*, 24(1):598.

Soya Nah, Jeeyun Oh, and Lucy Atkinson and. 2023. Interactive health messages work better for those who feel less in control: The role of external health locus of control and risk perception. *Journal of Health Communication*, 28(5).

Alexandra Olsacher, Celina Bade, Jan Ehlers, and Leonard Fehring. 2023. How to effectively communicate health information on social media depending on the audience's personality traits: An experimental study in the context of organ donation in germany. *Social Science & Medicine*, 335:116226.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2026. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *Preprint*, arXiv:2411.10109.

Rifaa Qadri, Anh Nhat Nhu, Swati Ramnath, Laura Yu Zheng, Raj Bhansali, Sylvette La Touche-Howard, Tracy Marie Zeeger, Tom Goldstein, and Ming Lin. 2025. Phorecast: Enabling ai understanding of public health outreach across populations. *Preprint*, arXiv:2510.02535.

Ann Nedime Nese Rende, Tolga Yilmaz, and Özgür Ulusoy. 2025. Negotiating comfort: Simulating personality-driven llm agents in shared residential social networks. *Preprint*, arXiv:2507.09657.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. LLMs simulate big5 personality traits: Further evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 83–87, St. Julians, Malta. Association for Computational Linguistics.

Antonio Taboada-Vázquez, Ruben Gonzalez-Rodriguez, Manuel Gandoy-Crego, and Miguel Clemente. 2021. Personality variables as predictors of health services consumption. *International Journal of Environmental Research and Public Health*, 18(10).

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar N. E. Kjell, Joao Sedoc, Margaret L. Kern, Ryan L. Boyd, Lyle Ungar, H. Andrew Schwartz, and Johannes C. Eichstaedt. 2025. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. *Preprint*, arXiv:2412.16882.

Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519.

Margaret A. Webb and June P. Tangney. 2024. Too good to be true: Bots and bad data from mechanical turk. *Perspectives on Psychological Science*, 19(6):887–890. PMID: 36343213.

Dan Weisman, Alanna Sugarman, Yue Ming Huang, Lillian Gelberg, Patricia A Ganz, and Warren Scott Comulada. 2025. Development of a gpt-4–powered virtual simulated patient and communication training platform for medical students to practice discussing abnormal mammogram results with patients: Multiphase study. *JMIR Form Res*, 9:e65670.

Supplementary Materials

A Model Prompting Methods

A.1 Attribute-Aware Response Prediction

This section describes the prompting methods for **Attribute-Aware VLM response prediction** described in Figure 3.

System Prompt. The system prompt describes the simulation scenario and role-playing task. It is used in both zero-shot and fine-tuned settings.

System Prompt

```
<|system|>
You are a person with unique demographic
and personality traits. During an online
study, you are shown a public health campaign
poster. Based on your background, you
naturally have thoughts, feelings,
and reactions to what you see.
```

Instruction Prompt. This prompt provides the individual's demographic and psychometric traits, along with a targeted survey question.

Instruction Prompt

```
<|user|>
You are: {demo_info}.
Your personality test shows you have
(min score = 0; max score = 5):
{persona_score}.
You also have {locus of control scores}.

You are now being shown a public health
campaign poster, followed by a survey
question designed to capture your thoughts,
feelings, and emotions responses to the image.

Please respond the survey question
authentically, as if you are completing
a real online survey. Your personality,
locus of control, and demographic traits
influence your reactions.

The health message: <IMG>
The question: {survey question}.

**CRITICAL INSTRUCTIONS
- FOLLOW THESE EXACTLY:**
- **BE REALISTIC and HUMAN-LIKE.
    Only answer the questions. Imagine
    you're quickly filling out a survey.**
- Answer with **ONLY the valid number**
    to realistically express your
    emotions/feeling.
- Only provide the answer.
  DO NOT REPEAT THE QUESTION
  NOR PROVIDE REASONING.
```

A.2 Likert-scale to Natural Language Conversion

We applied contextual framing and few-shot prompting to convert predicted Likert responses to natural language simulation.

System Prompt

```
<|system|> You are a helpful assistant.
```

Instruction Prompt

```
<|user|>
Summarize following survey responses into a
short, natural paragraph that captures your
overall sentiment, motivation, and thinking.
Write as if paraphrasing what a person may say
in conversation. Adjust style based on your
demographic/personality traits. Do NOT repeat
numeric scores.
Preserve polarity:
low scores → low concern/motivation/openness;
high scores → high concern/motivation/openness
If answers are mixed (e.g., believes something
is harmful but isn't personally moved), reflect
that nuance explicitly. Keep to 1-5 sentences.

**STRICTLY FOLLOW THESE RULES:**
- Infer direction from each item's Scale
  description (e.g., 1-9: higher = more; 0-6:
   higher = more).
- Use calibrated wording: 1-2 = very low,
   3-4 = low, 5 = moderate, 6-7 = high,
  8-9 = very high; for 0-6: 0-1 = not/slight,
   2-3 = somewhat, 4-5 = high, 6 = very.
- VERY IMPORTANT: provide ONLY the final
  summarized response, without anything else!
- Start the response with 'This message...' or
    'From this message...' or 'Through the
    message...' or 'After seeing this ...'
- The response MUST have a consistent health
  topic: {health_topic}. Ground each sentence
   to the impact of campaign message.
- Never invert sentiment. Prefer hedged phrases
   (e.g., “not particularly,” “only somewhat,”
    “very open,” “not open at all”).

**Example input 1:**
The message makes me more concerned about the
health risks of poor eating habits - Scale: 1-9.
Your answer: 9
The message motivates me to make healthy eating
choices - Scale: 1-9. Your answer: 9
In your opinion, how harmful is neglecting
proper nutrition and weight management to
your overall health?
- Scale: 0-6. Your answer: 5
How open are you to adopting healthier eating
habits and lifestyle changes? - Scale: 1-9.
Your answer: 9

**Example output 1:**
This message really heightened my awareness of
how unhealthy eating can be. The content in the
message strongly motivates me to make better
choices, and I feel very ready to follow .
```

B Multi-Disciplinary Collaboration

This project was conducted as multi-disciplinary collaboration between Human-Centric AI researchers and School of Public Health (SPH) faculty and graduate students, who provided the driving insights of this critical need for public health, required domain expertise and educational perspectives throughout the entire study. Our Public Health team and HCI expert guided the design and review of survey instruments used in dataset construction to ensure that the questions and responses appropriately captured the reactions to health campaign messages.

SPH faculty also offered iterative feedback on model development and system design, to help align the response simulation framework with realistic classroom needs and instructional objectives. Their perspectives, along with HCI guidance, informed the selection of demographic and psychometric attributes, the interpretation of simulated responses, and the design of evaluation protocols to assess educational guidelines and realism, with direct access to an undergraduate public-health course for the reported in-class pilot study, enabling evaluation of this novel system in an authentic classroom setting. The resulting framework allowed us to assess how students interact with simulated communities in practice and to identify functionalities most relevant to experiential learning that typically would require time-consuming and laborious field studies. Via attributed-aligned LVLM to simulate diverse community responses, students were able to obtain real-time feedback emulating individuals of varying demographic and psychometric traits otherwise difficult to achieve in class.

C Model Training and Inference Hyperparameters

For all experiments, we trained each model using a single Nvidia A5000 24G GPU. For efficient training, the maximum input sequence length (context window) is set to 2048 with LoRA rank $r = 8$. We froze the vision encoders and fine-tuned only the language layers, including the attention layers and MLP layers. `paged_adamw_8bit`, an 8-bit AdamW variant with paged memory offloading, was used as the optimizer for all experiments. All other hyperparameters, such as learning rate and optimizer, were set to Unsloth's default values.

D Data Quality Assurance/Screening and Response Distribution

Data Quality Assurance and Participant Screening. To ensure the highest quality of the response dataset for trait alignment training and evaluation, we applied multiple screening steps to the original responses. ***First***, we removed bot-like entries flagged by an automated filter (inspired by Webb and Tangney (2024)) that detected repetitive, nonsensical, or unnatural text. ***Second***, we applied effort-based filtering using recorded survey duration thresholds ($\geq$600 seconds (10 minutes) and $\leq$7,200 seconds (2 hours)), reducing nearly 1,950 subjects to 1,457. ***Third***, we excluded participants who left all demographic or personality trait questions unanswered, further reducing the number of valid respondents to 1,119. ***Finally***, we performed IP-based consistency checks, removing cases where the same IP address was linked to conflicting or non-unique trait data, which we interpreted as low-effort submissions motivated primarily by survey rewards rather than faithful participation. This last step filtered an additional 108 subjects, yielding a final dataset of 1,011 high-quality, unique participants. While shared networks are possible, we opted for this cautious approach to ensure data quality. In summary, these screening steps filtered nearly 950 potentially low-quality participants (nearly 50% of the original sample), ensuring only high-quality responses were retained for trait alignment training and evaluation.

E Survey Participant, Data Consent, and Board Approval

Ethics Board Approval. The survey and the response data collection were reviewed and approved by the Institutional Review Board (IRB). We used the response data strictly for academic research, which is the scope of this paper.

Data Consent. Participants were informed in the consent section that the data would be anonymized before use. The data consent information given to each survey participant was "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 the institution 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."

Pilot Test Participants. With permission from public health lecturers, we conducted pilot studies in upper-level courses on Media Communication and Digital Tools in Public Health, with students as the targeted end users. Students were not compensated for participation. Students' identities and evaluations were kept anonymous throughout the pilot study. We did not collect or associate identifying information with students' use of CLAAS or their Google Forms feedback. Students were informed that their anonymous responses would be used for academic research purposes.

F Sample Response and Trait Alignment Dataset Documentations

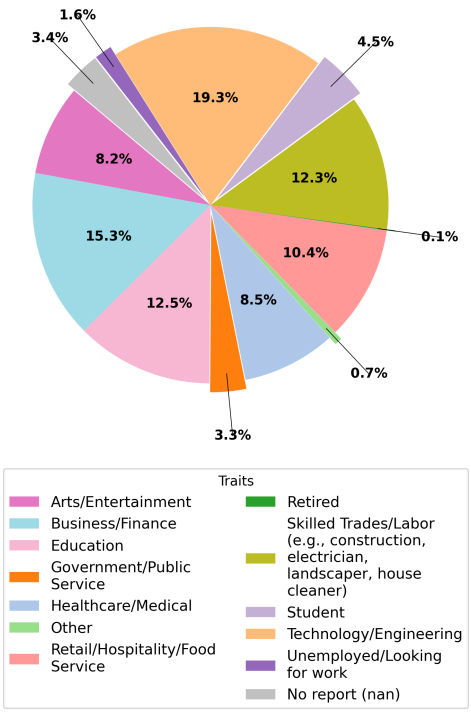


Figure 10: Dataset composition by ***current profession*** (valid participants after screening).

Dataset Statistics. The raw dataset has nearly 2,000 subjects. However, after applying our screening process to ensure data quality, the number of subjects included in this study was reduced to 1,011. Each participant is shown and asked to share their responses to 2-3 posters on different health topics, each of which consists of 13 questions measuring their sentimental, emotional, and behavioral reactions to each poster. Each participant was shown posters on different health topics and asked to answer 13 questions/poster. In total, the dataset has 33,683 responses (1,011 subjects×2.56 posters×13 questions/poster).

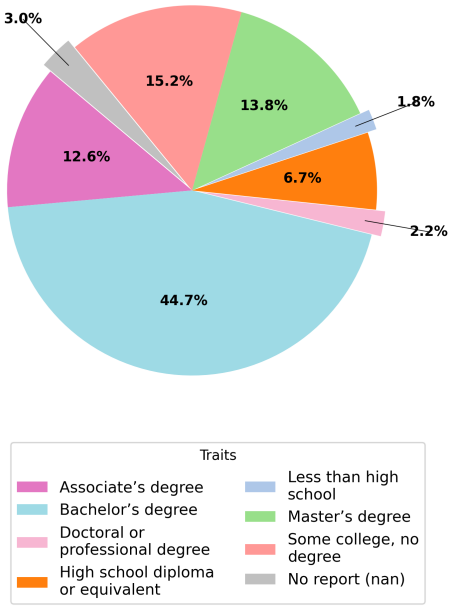


Figure 11: Dataset composition by ***highest education level*** (valid participants after screening).

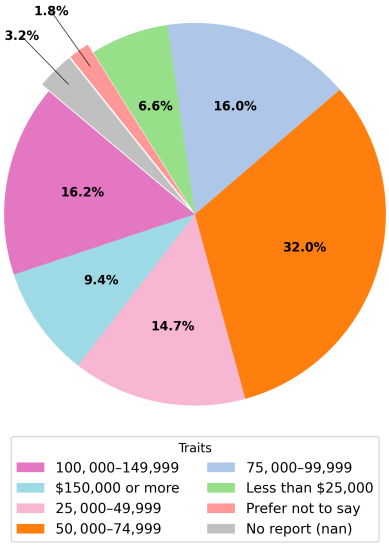


Figure 12: Dataset composition by ***annual household income*** (valid participants after screening).

We further provide some demographic composition details in Figures 10, 11, 12, 13, 14, 15, 16. Figures 6, 7, 8, 20, 19, 21, 22, 23, 24, 25, 26 show additional insights into response distributions across trait groups, with ground-truth responses shown as white bars. Although many respondents gave

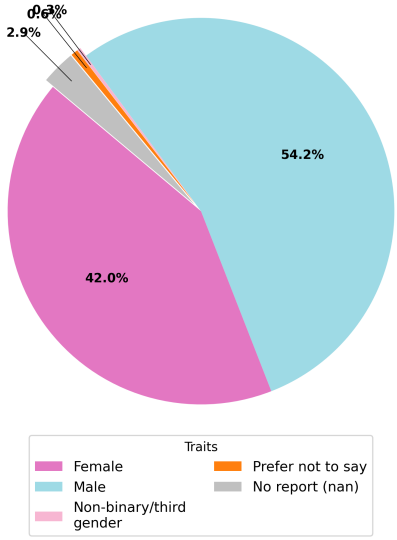


Figure 13: Dataset composition by ***gender*** (valid participants after screening).

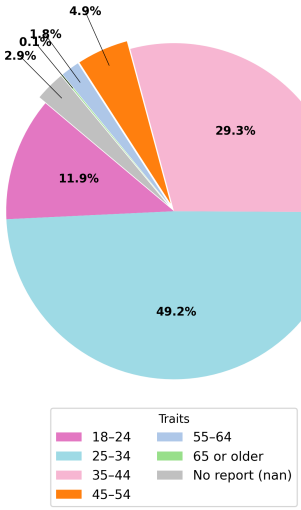


Figure 14: Dataset composition by ***age*** (valid participants after screening).

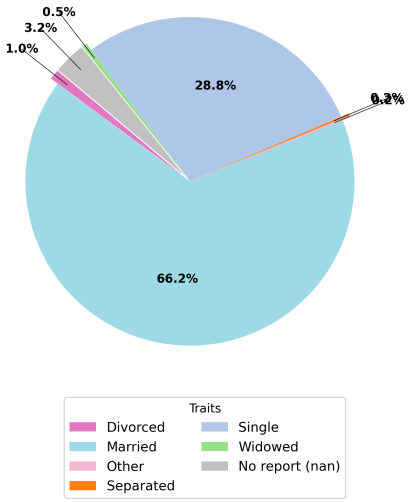


Figure 15: Dataset composition by ***marital status*** (valid participants after screening).

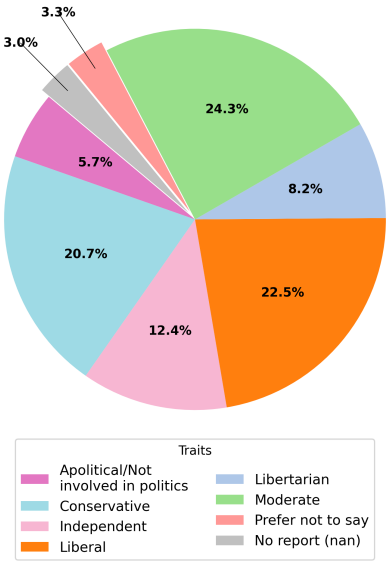


Figure 16: Dataset composition by ***political affiliation*** (valid participants after screening).

generally positive ratings, the distributions varied between trait groups. This pattern is expected, as people often respond mildly positively unless they feel strongly in agreement or opposition to a message. Figure 7 and Figure 25 illustrate these distributional differences. Notably, Figure 25 shows that the *High Agreeableness* group reports nearly twice as many "very positive" responses (48.51%) as the *Moderate Agreeableness* group (26.47%), consistent with the expectation that highly agreeable individuals respond more positively to health messaging.

G Additional Results and Analysis

G.1 Statistical Significance Testing of Performance Gains

To verify that gains from attribute alignment are statistically significant, we compare attribute-aligned models against the strongest baselines, i.e., zero-shot VLMs, using **both** exact one-sided McNemar test (McNemar, 1947) and cluster-robust paired bootstrap test (Efron and Tibshirani, 1994) over paired correctness outcomes. For each tolerance level $tol \in \{0, 1, 2\}$, a prediction is labeled correct if $|\hat{y}_i - y_i| \leq tol$. Let n_{10} denote the number of samples where the attribute-aligned model is correct and the baseline is incorrect, and let n_{01} denote

the opposite:

$$n_{10} = \sum_i \mathbf{1}(c_i^{ours} = 1, c_i^{baseline} = 0),$$
$$n_{01} = \sum_i \mathbf{1}(c_i^{ours} = 0, c_i^{baseline} = 1).$$

Under the null hypothesis, both models are equally likely to win on discordant samples. We compute the one-sided exact McNemar p-value as

$$p = P(X \geq n_{10}), \quad X \sim \text{Binomial}(n_{10} + n_{01}, 0.5),$$

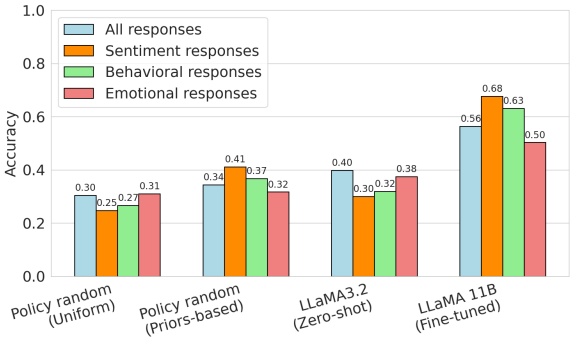
testing whether the attribute-aligned model wins more often than the baseline. Additionally, to account for repeated-measures dependency and intra-participant correlation across the survey, we perform a cluster-robust paired bootstrap ($B = 10{,}000$ resamples clustered at the participant level and 95% CI) to compute cluster-adjusted empirical p-values and 95% confidence intervals. The tests are applied separately to exact match, ± 1, and ± 2 tolerance accuracy across all VLM backbones and held-out communication strategies. We consider improvements significant at $p < 0.0001$. Under both statistical tests, ***performance gains are statistically significant across all VLM backbones (Gemma, LLaMA, Qwen), metrics, and held-out strategies.***

G.2 Evaluation of Response Prediction for `Llama-3.2:11B` and `Qwen2.5-VL:7B`

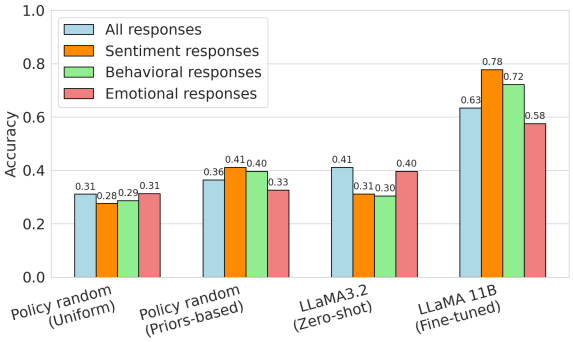
Figure 17 provides additional results to Figure 5 to show the effect of response alignment training for `Llama-3.2-Vision` model. We also further include similar investigation of the `Qwen2.5-VL:7B` model in Figure 18. Note that in these figures, the `Policy random (Priors-based)` differs from the `Policy random (Uniform)` in that, instead of uniformly sampling Likert-scale responses from 1 to 9, it adjusts the sampling distribution according to the prior probability distribution of each question type (sentimental, behavioral, or emotional). This creates a stronger random baseline for a more comprehensive comparison. Table 2 describes the prior Likert response score distributions to different question types. As expected, although the prior-based random baselines perform better than the uniform ones, our trait-aligned VLMs still outperform these stronger baselines.

G.3 Group-conditioned Response Distribution

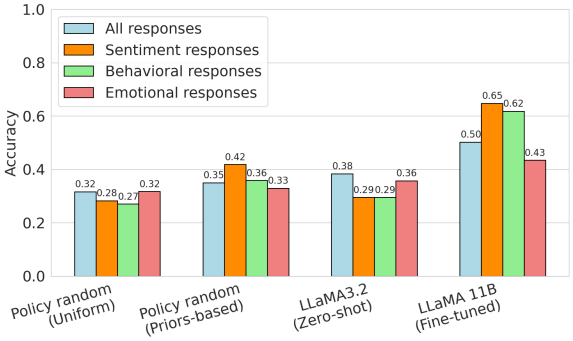
In this section, we report additional response distribution across representative demographic and psychometric groups in Figure 19 to Figure 42.



(a) Informational/Educational Health Messages

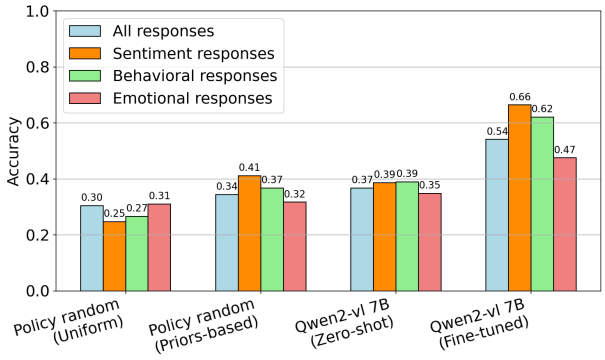


(b) Self-Efficacy Health Messages

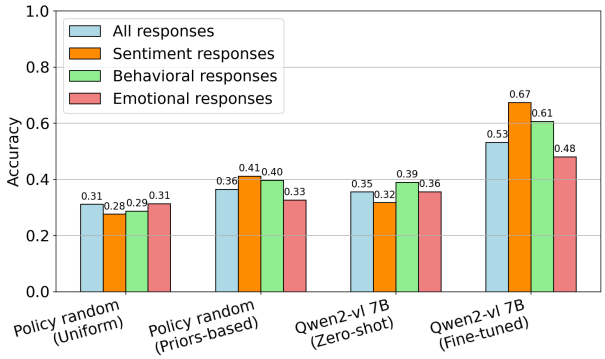


(c) Threatening/Fear-driven Health Messages

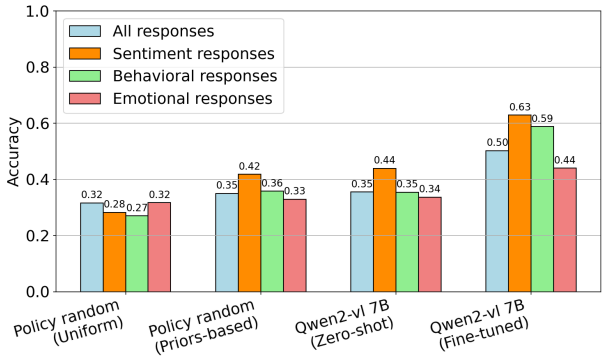
Figure 17: Llama-3.2-Vision Models' response prediction accuracy on various unseen communication strategies. As human sentimental and emotional survey scores are inherently imprecise, we report accuracy using a ± 1 error tolerance w.r.t. ground-truth scores to flexibly and practically capture overall trends and response patterns. The response alignment training improves the accuracy substantially across all response types (**37%–50%** for sentiment, **35%–43%** for behavior, and **11%–27%** for emotion) compared to the random baseline. ***These gains are statistically significant at*** $p \leq 0.0001$ under both exact one-side McNemar and cluster-robust pair bootstrap statistical tests ($B = 10{,}000$ and 95% CI).



(a) Informational/Educational Health Messages



(b) Self-Efficacy Health Messages



(c) Threatening/Fear-driven Health Messages

Figure 18: Qwen2.5-VL 7B Models' response prediction accuracy on various unseen communication strategies. As human sentimental and emotional survey scores are inherently imprecise, we report accuracy using a ± 1 error tolerance w.r.t. ground-truth scores to flexibly and practically capture overall trends and response patterns. The response alignment training improves the accuracy substantially across all response types (**34%–39%** for sentiment, **29%–34%** for behavior, and **12%–15%** for emotion) compared to the random baseline. ***These gains are statistically significant at*** $p \leq 0.0001$ under both exact one-side McNemar and cluster-robust pair bootstrap statistical tests ($B = 10{,}000$ and 95% CI).

Table 2: Prior Likert response score distributions for valid (screened) samples.

Likert Score	Sentiment	Behavior	Emotion
1	0.6 %	1.3 %	17 %
2	1.1 %	3.8 %	10.6 %
3	2.0 %	5.2 %	9.0 %
4	4.7 %	5.0 %	7.5 %
5	12.2 %	11.9 %	13.0 %
6	15.1 %	11.5 %	12.1 %
7	17.5 %	17.1 %	13.8 %
8	17.7 %	15.9 %	8.6 %
9	29.1 %	28.3 %	8.4 %
$\sum$	100 %	100 %	100 %

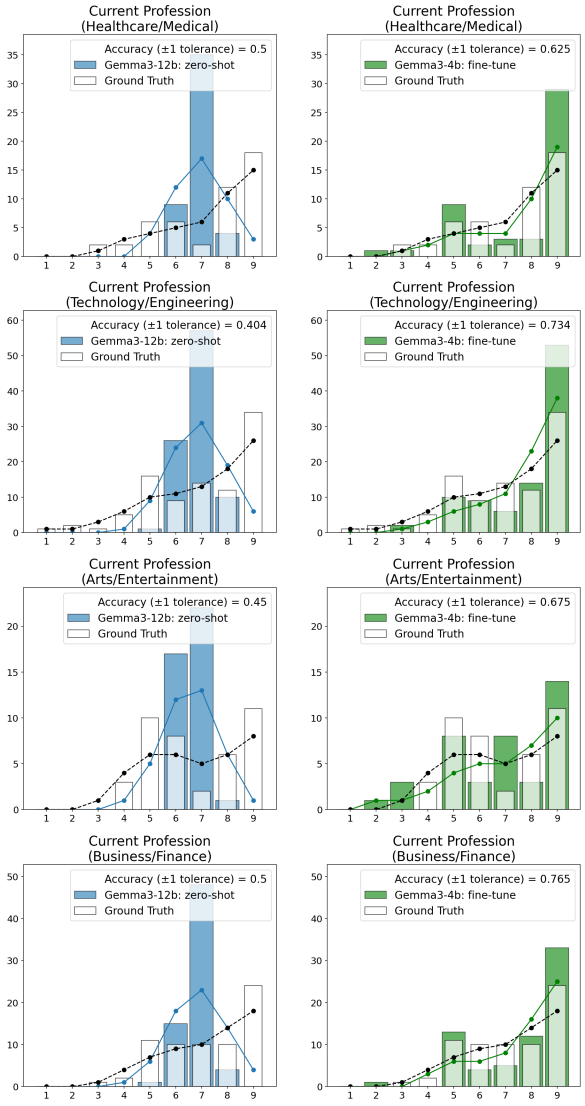


Figure 19: Sentimental response distributions of **Gemma3** models to unseen **Self-Efficacy** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

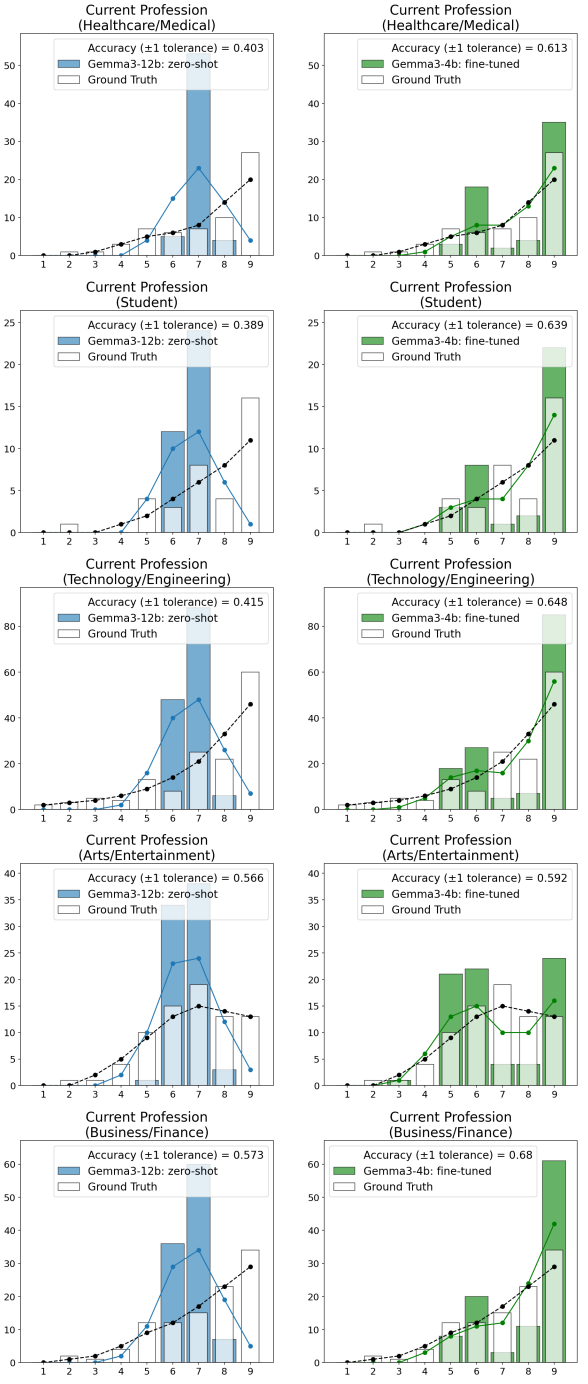


Figure 20: Sentimental response distributions of **Gemma3** models to unseen **Informational / Educational / Neutral** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

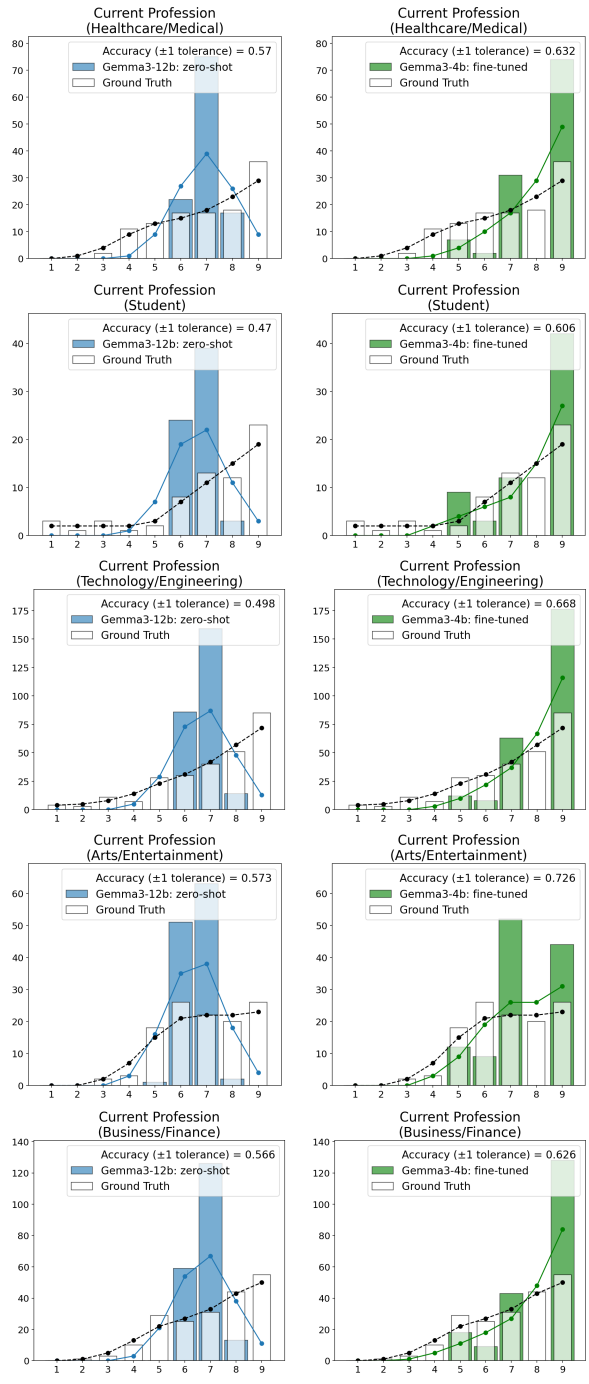
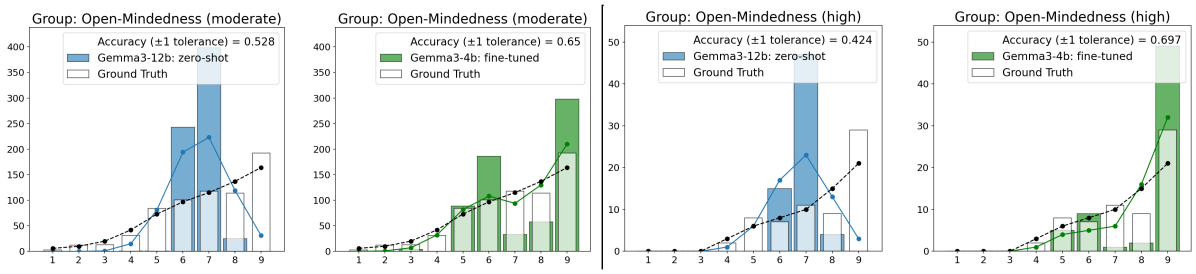
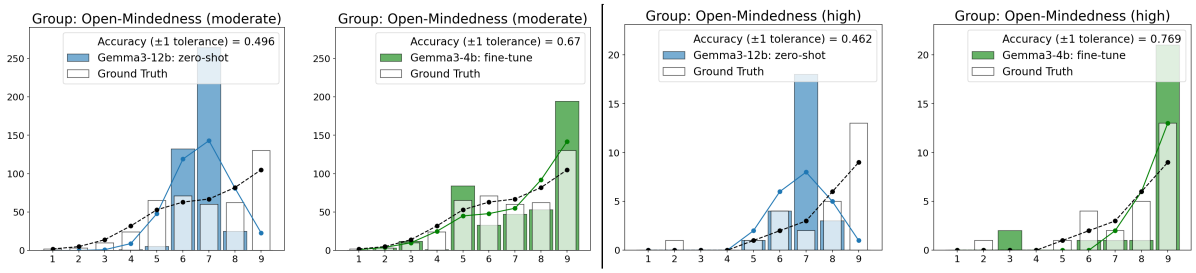


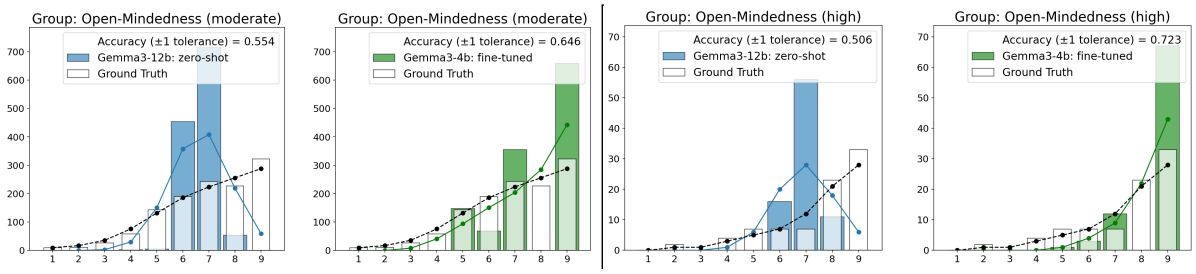
Figure 21: Sentimental response distributions of **Gemma3** models to unseen **Threatening/Fear-driven** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

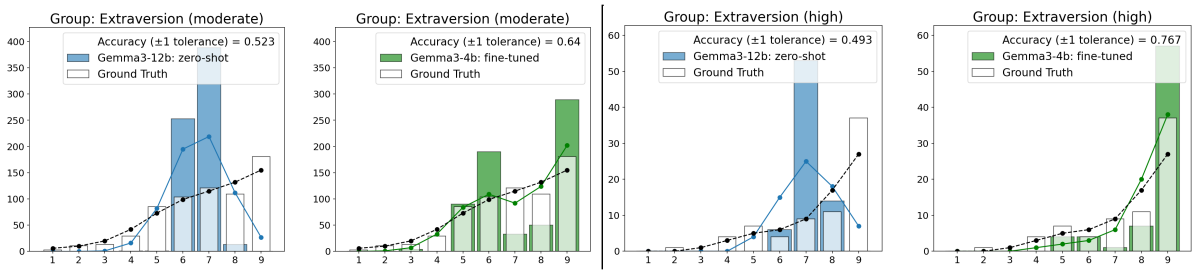


(b) Self-Efficacy Health Messages

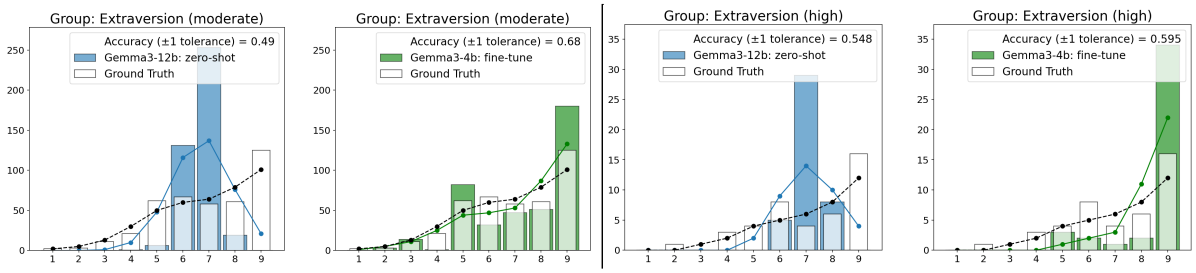


(c) Threatening/Fear-driven Health Messages

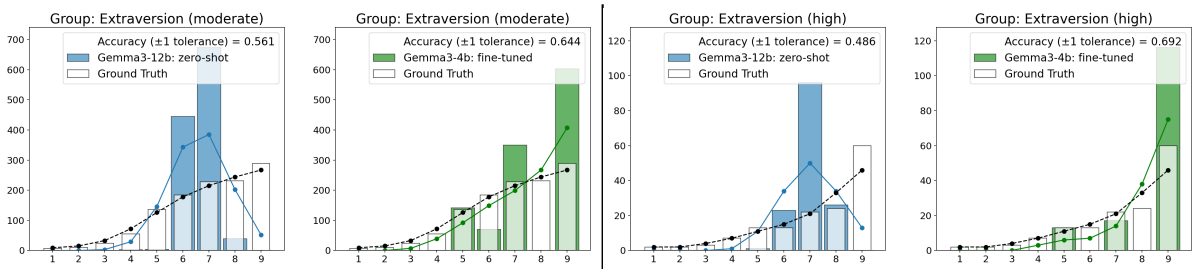
Figure 22: Sentimental response distributions of **Gemma3** models to unseen communication strategies across **"Open-Mindedness" personality groups** (*moderate, high*). "*Low* Open-Mindedness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

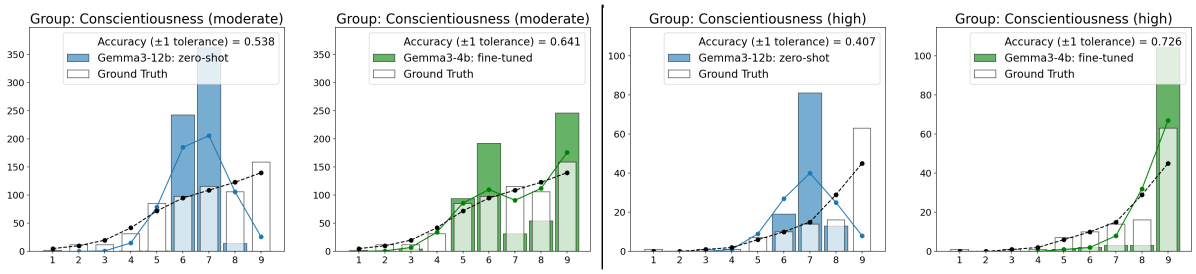


(b) Self-Efficacy Health Messages

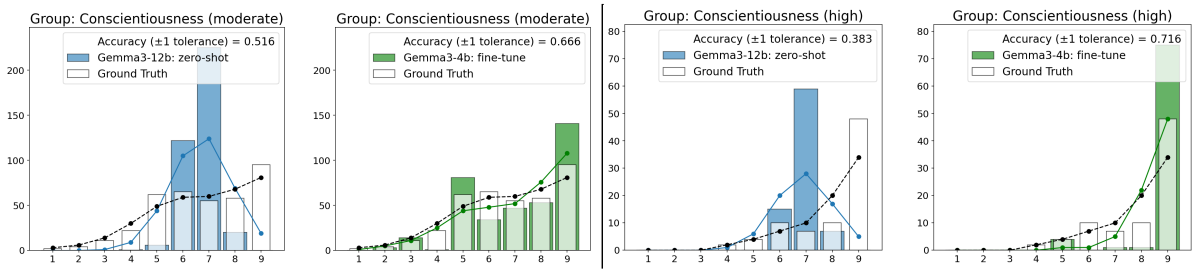


(c) Threatening/Fear-driven Health Messages

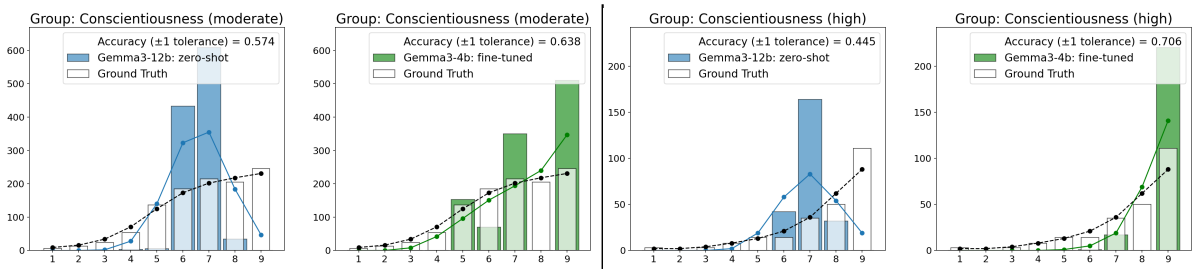
Figure 23: Sentimental response distributions of **Gemma3** models to unseen communication strategies across **"Extraversion" personality groups** (*moderate, high*). "*Low* Extraversion" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

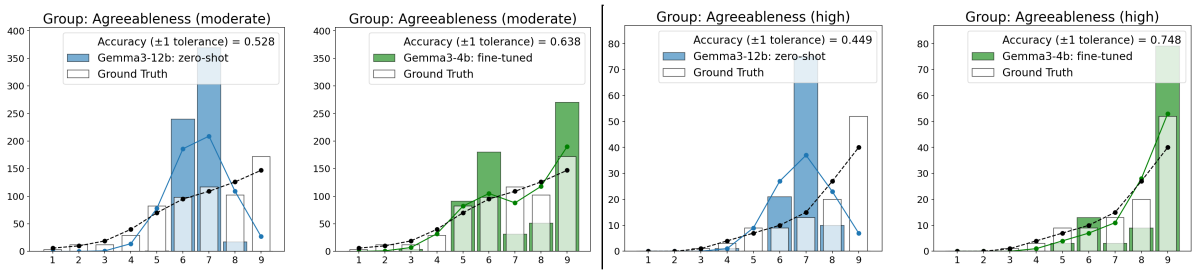


(b) Self-Efficacy Health Messages

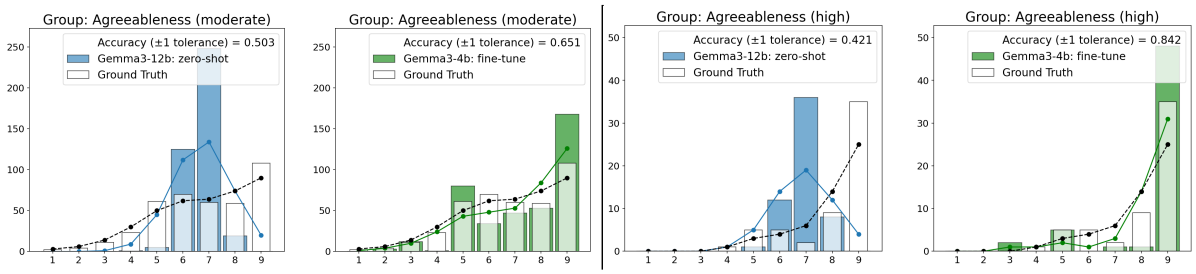


(c) Threatening/Fear-driven Health Messages

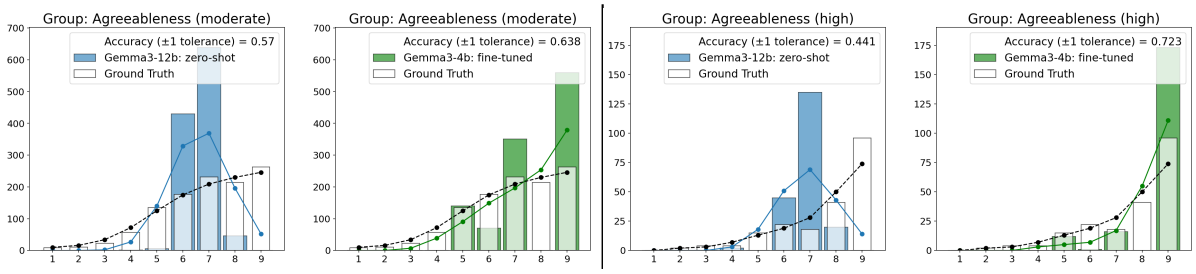
Figure 24: Sentimental response distributions of **Gemma3** models to unseen communication strategies across **"Conscientiousness" personality groups** (*moderate, high*). "*Low* Conscientiousness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

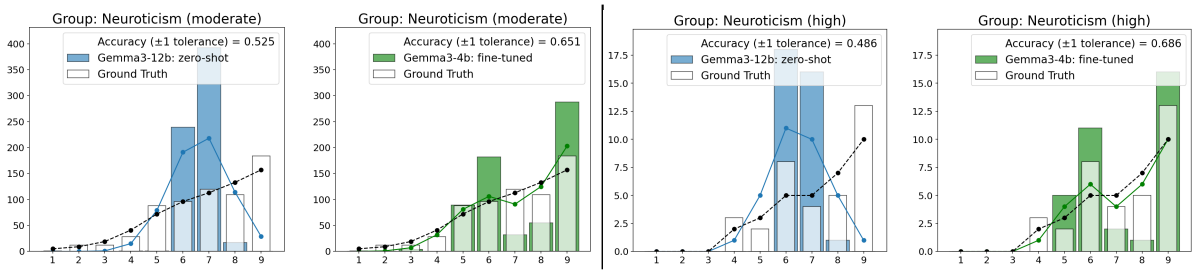


(b) Self-Efficacy Health Messages

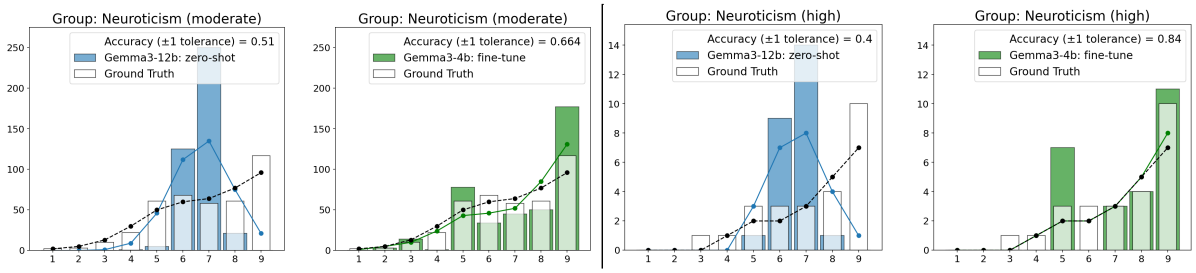


(c) Threatening/Fear-driven Health Messages

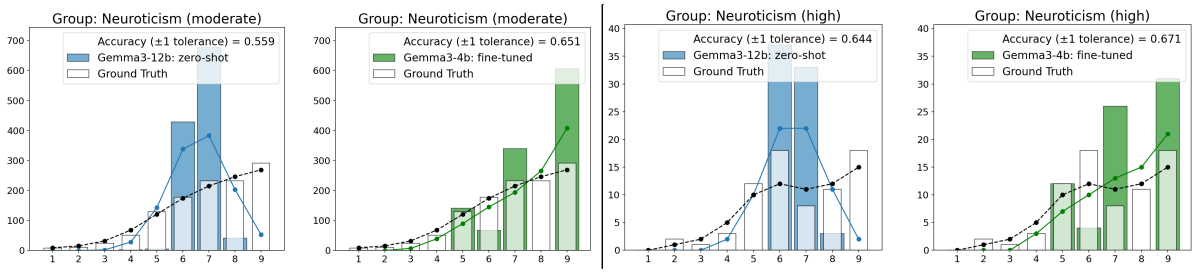
Figure 25: Sentimental response distributions of **Gemma3** models to unseen communication strategies across **"Agreeableness" personality groups** (*moderate, high*). "*Low* Agreeableness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages



(b) Self-Efficacy Health Messages



(c) Threatening/Fear-driven Health Messages

Figure 26: Sentimental response distributions of **Gemma3** models to unseen communication strategies across **"Neurocitism" personality groups** (*moderate, high*). Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

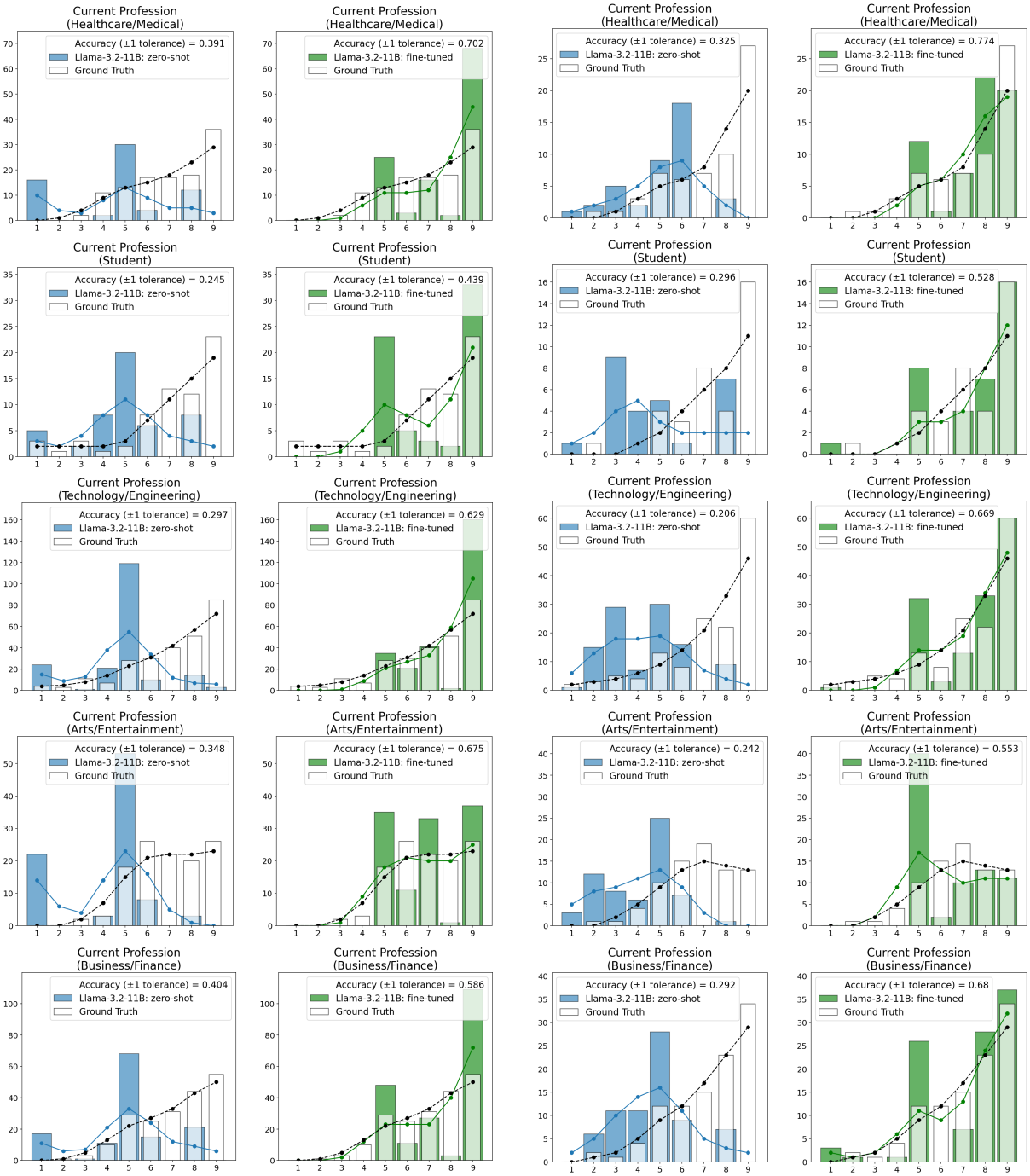


Figure 27: Sentimental response distributions of **Llama3.2** models to unseen **Threatening/Fear-driven** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

Figure 28: Sentimental response distributions of **Llama3.2** models to unseen **Informational / Educational / Neutral** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

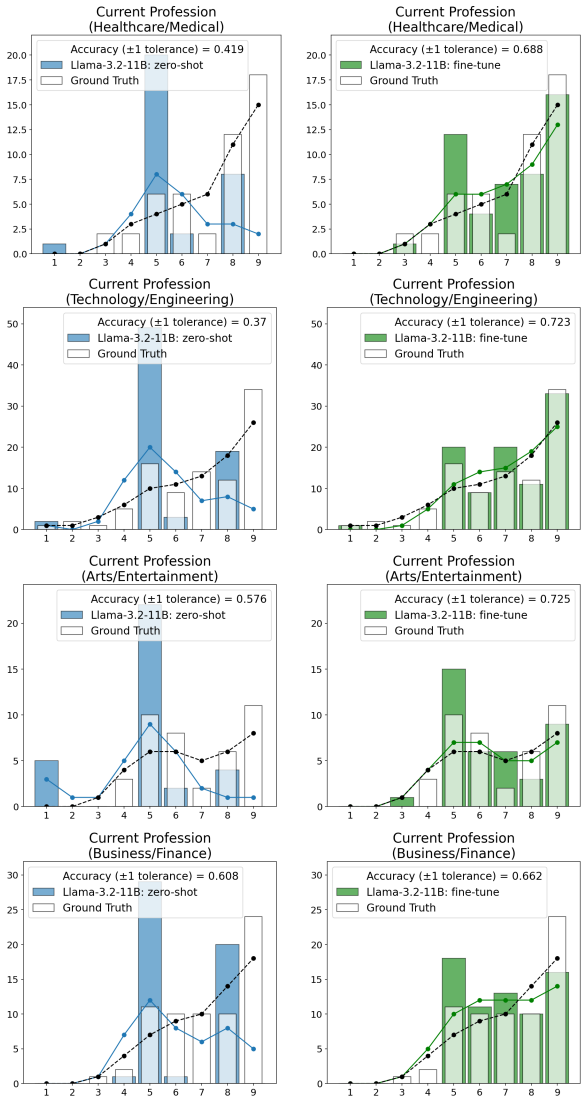
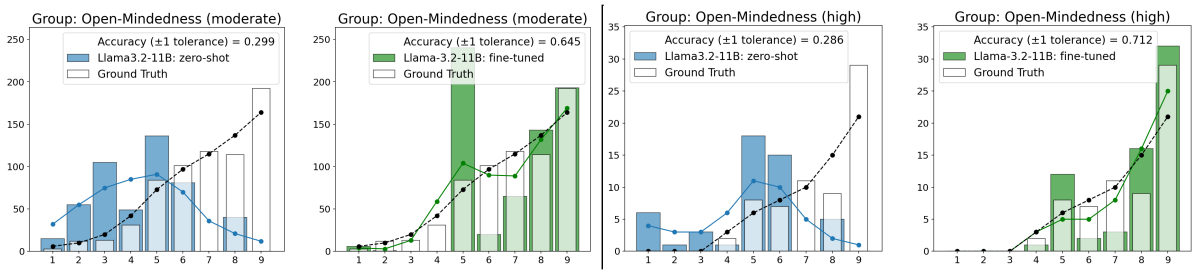
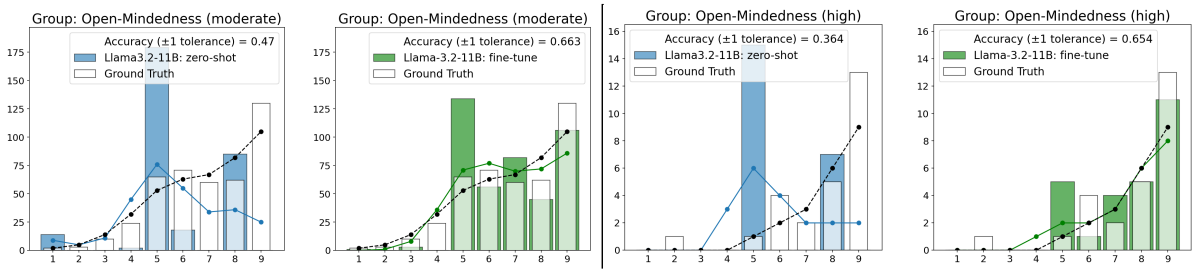


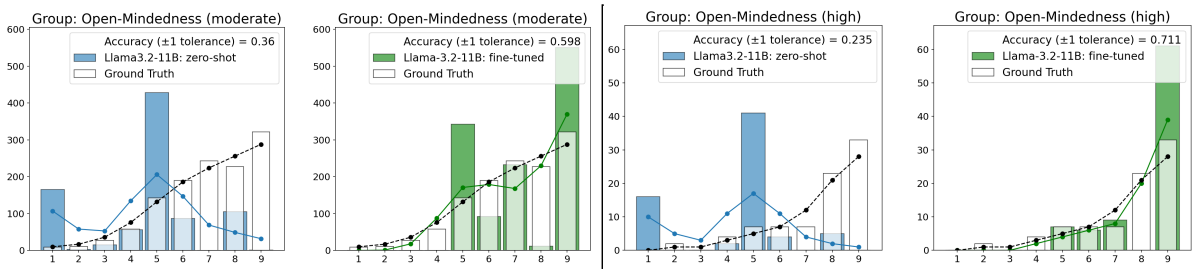
Figure 29: Sentimental response distributions of **Llama3.2** models to unseen **Self-Efficacy** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

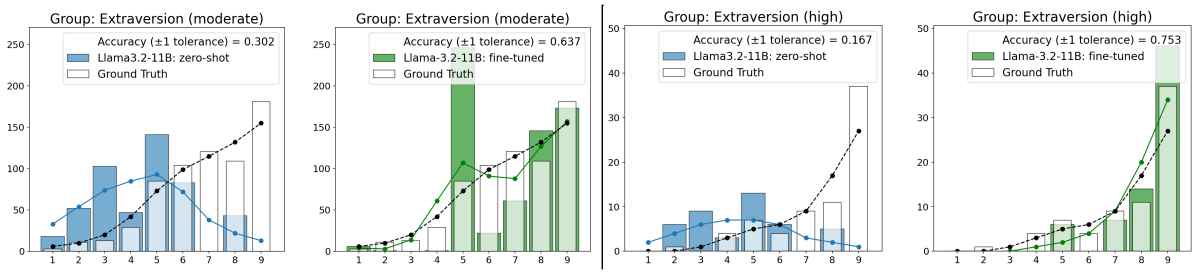


(b) Self-Efficacy Health Messages

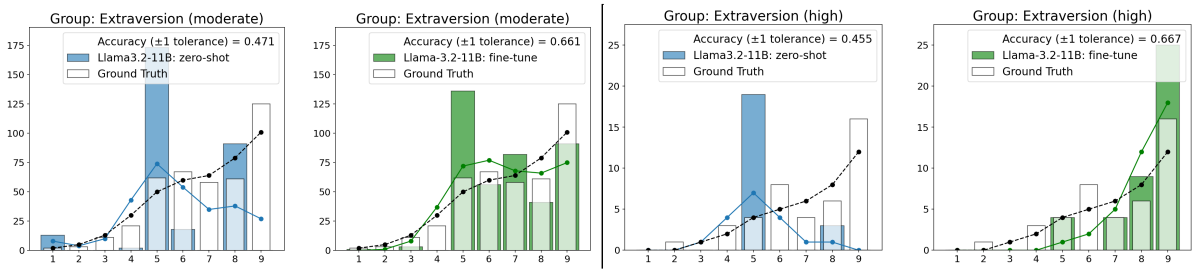


(c) Threatening/Fear-driven Health Messages

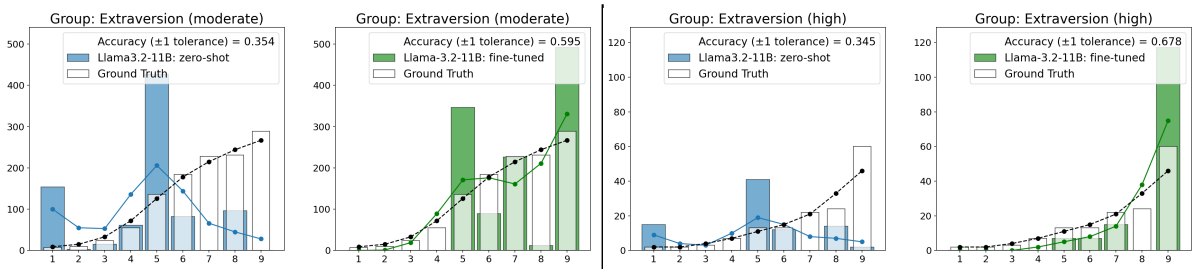
Figure 30: Sentimental response distributions of **Llama3.2** models to unseen communication strategies across **"Open-Mindedness" personality groups** (*moderate, high*). "*Low* Open-Mindedness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

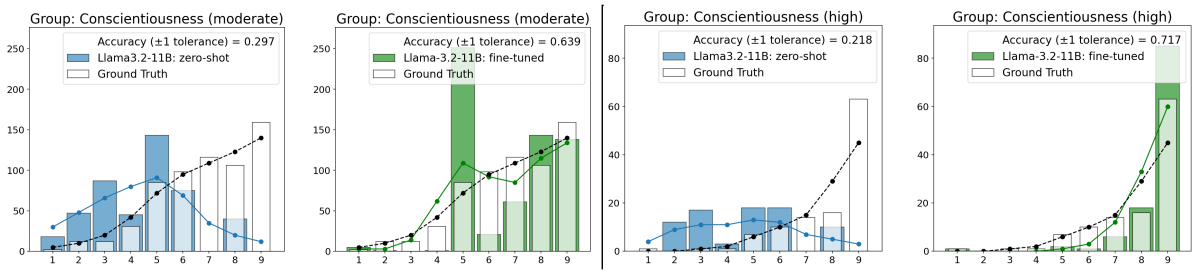


(b) Self-Efficacy Health Messages

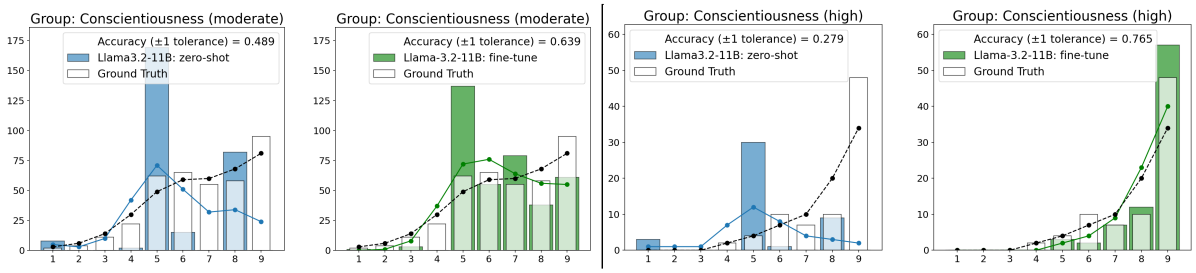


(c) Threatening/Fear-driven Health Messages

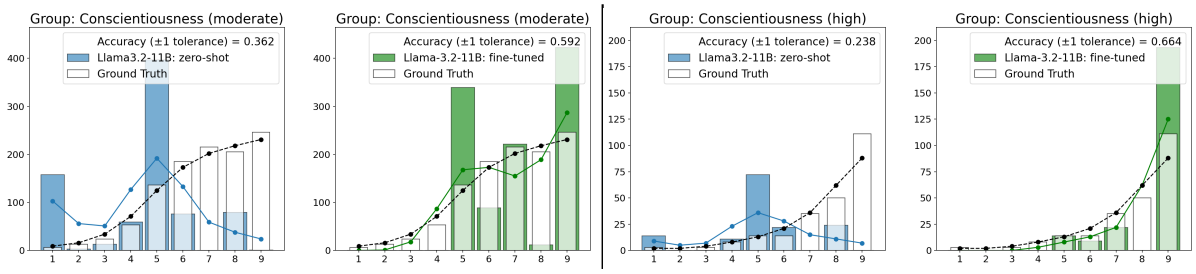
Figure 31: Sentimental response distributions of **Llama3.2** models to unseen communication strategies across **"Extraversion" personality groups** (*moderate, high*). "*Low* Extraversion" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

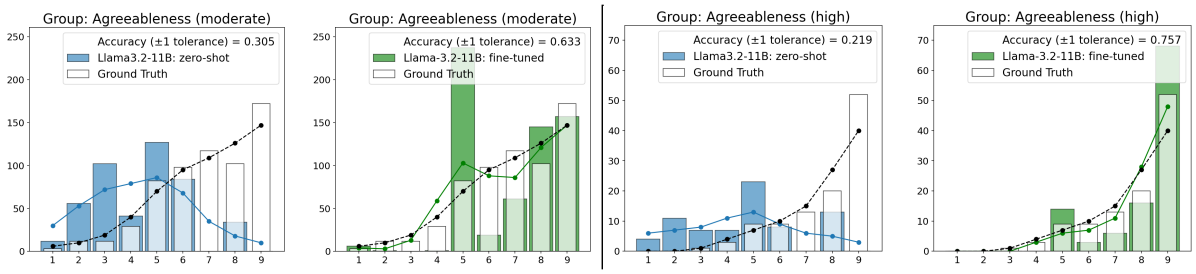


(b) Self-Efficacy Health Messages

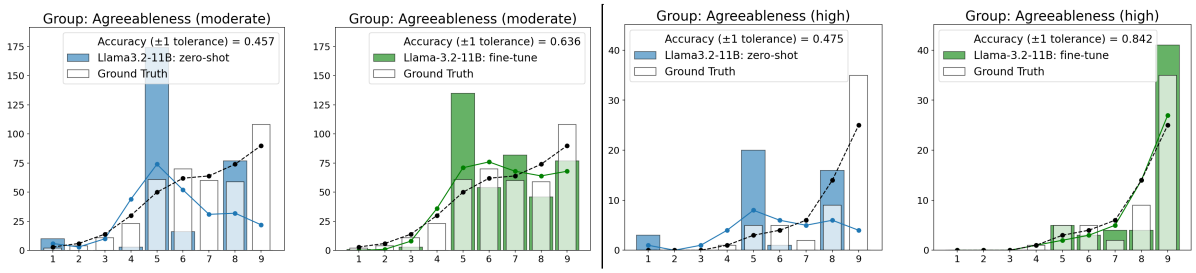


(c) Threatening/Fear-driven Health Messages

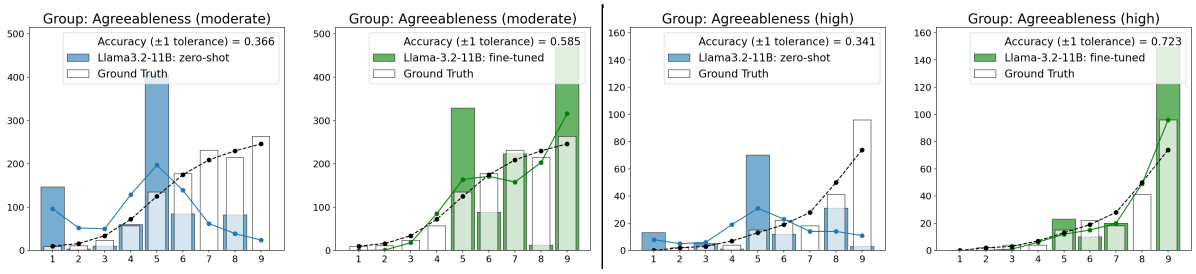
Figure 32: Sentimental response distributions of **Llama3.2** models to unseen communication strategies across **"Conscientiousness" personality groups** (*moderate, high*). "*Low* Conscientiousness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

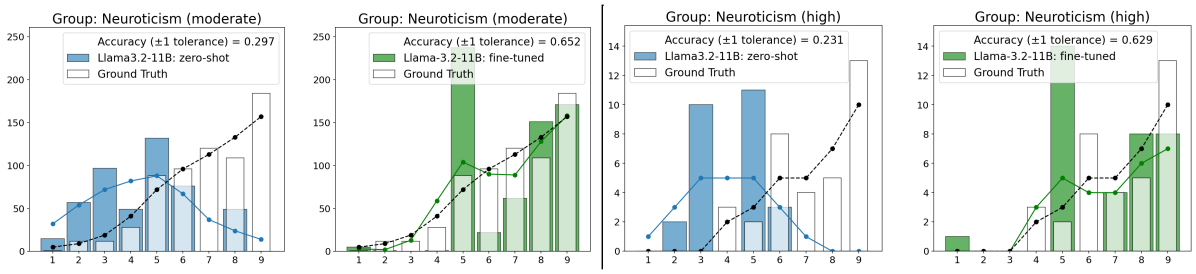


(b) Self-Efficacy Health Messages

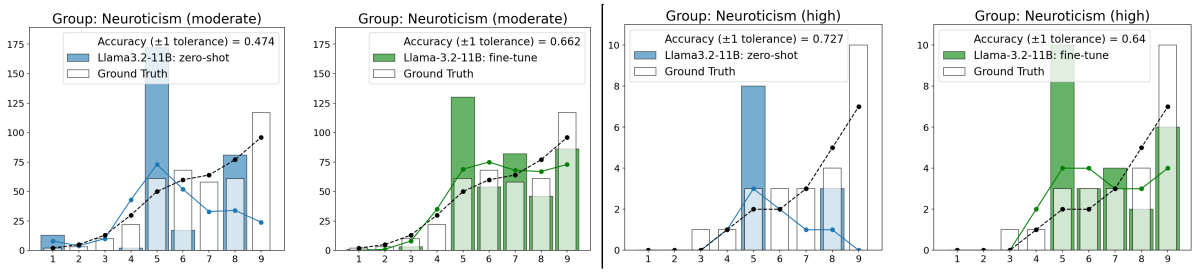


(c) Threatening/Fear-driven Health Messages

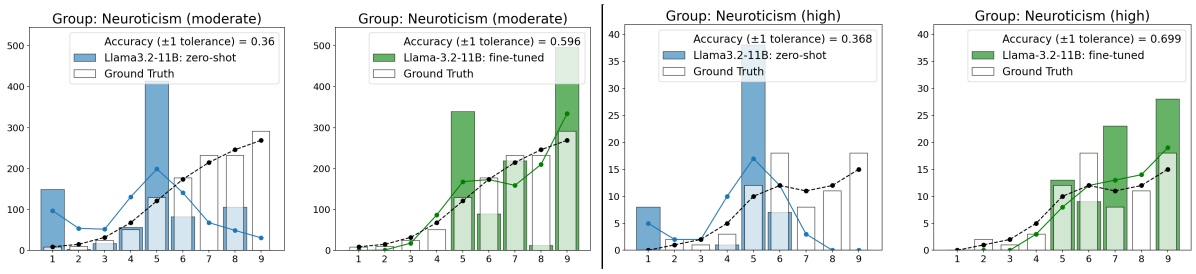
Figure 33: Sentimental response distributions of **Llama3.2** models to unseen communication strategies across **"Agreeableness" personality groups** (*moderate, high*). "*Low* Agreeableness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages



(b) Self-Efficacy Health Messages



(c) Threatening/Fear-driven Health Messages

Figure 34: Sentimental response distributions of **Llama3.2** models to unseen communication strategies across **"Neurocitism" personality groups** (*moderate, high*). Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

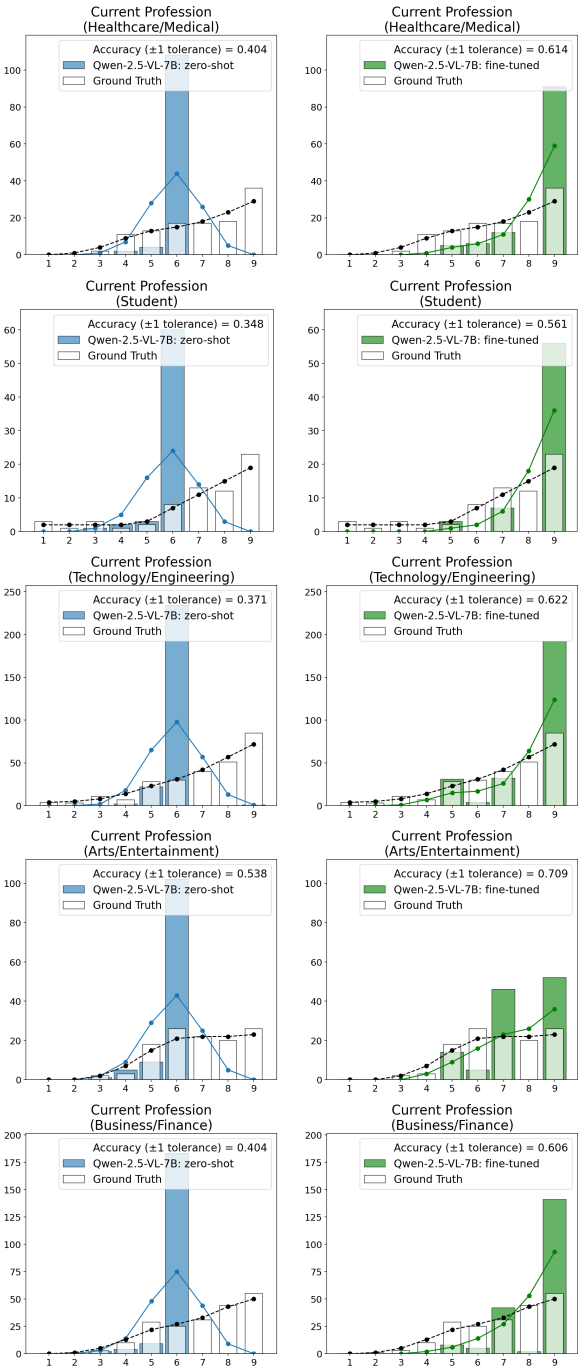


Figure 35: Sentimental response distributions of **Qwen2.5-VL** models to unseen **Threatening/Fear-driven** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

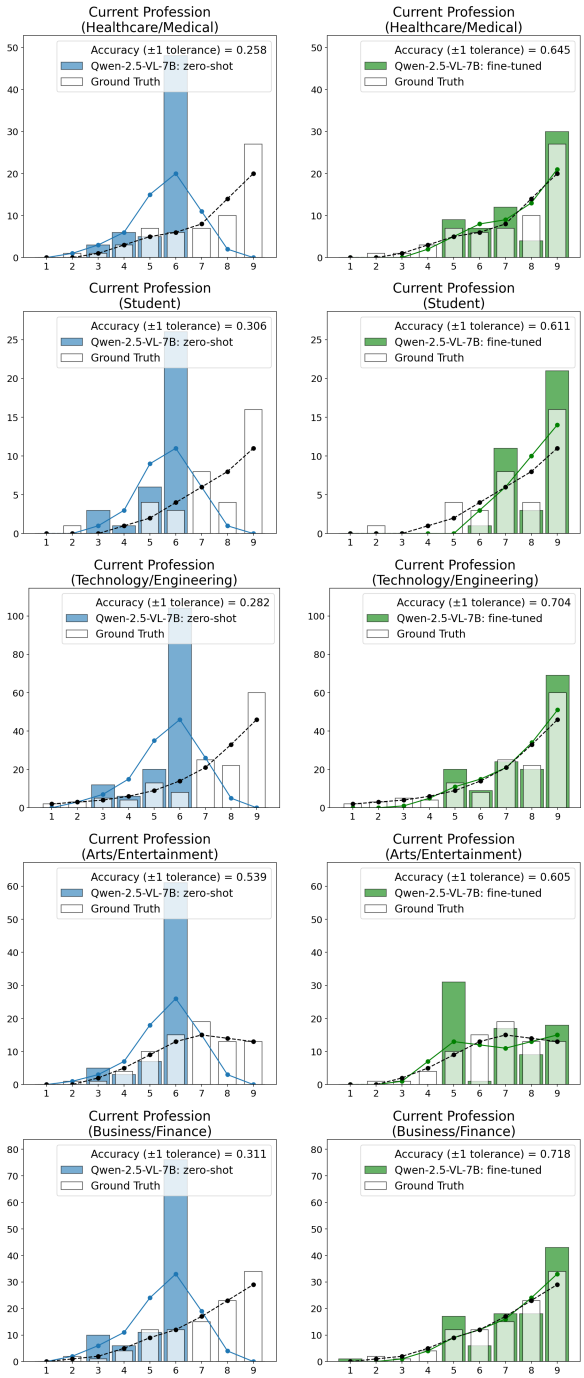


Figure 36: Sentimental response distributions of **Qwen2.5-VL** models to unseen **Informational / Educational / Neutral** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

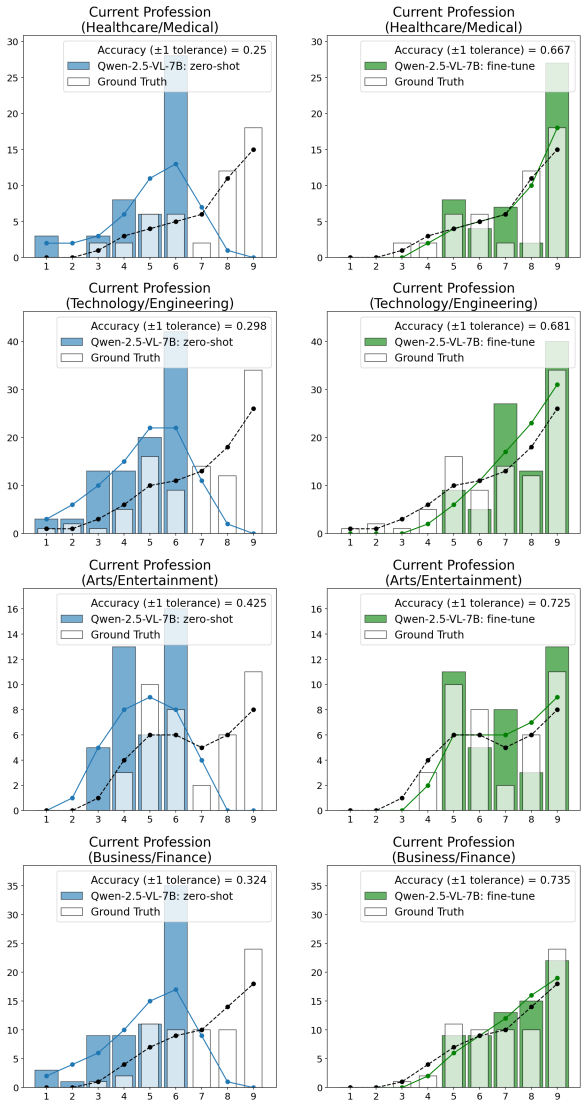
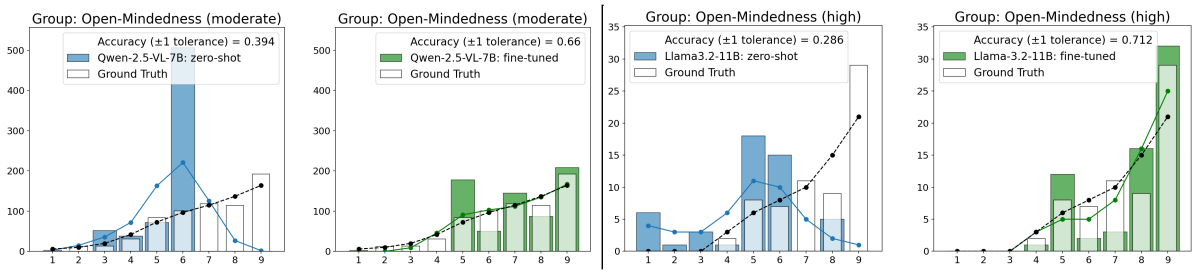
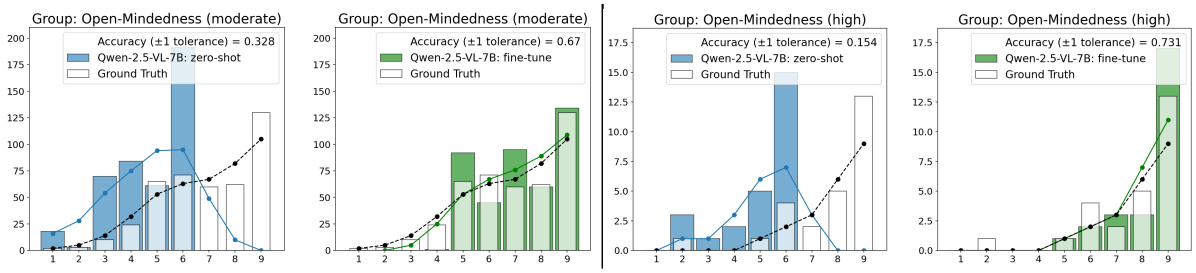


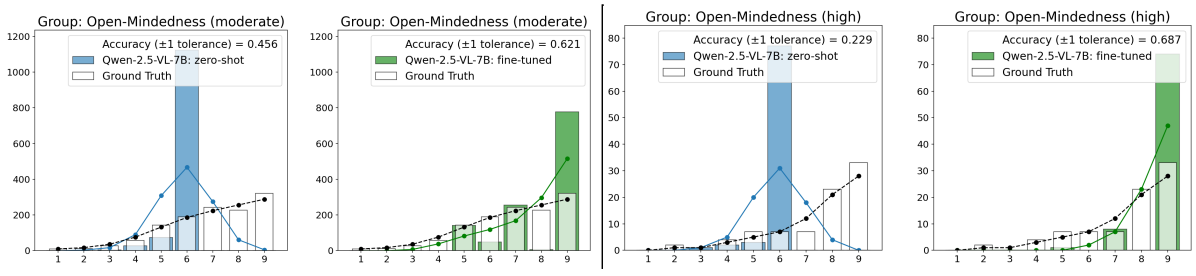
Figure 37: Sentimental response distributions of **Qwen2.5-VL** models to unseen **Self-Efficacy** messages across **"Current Profession" groups**. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

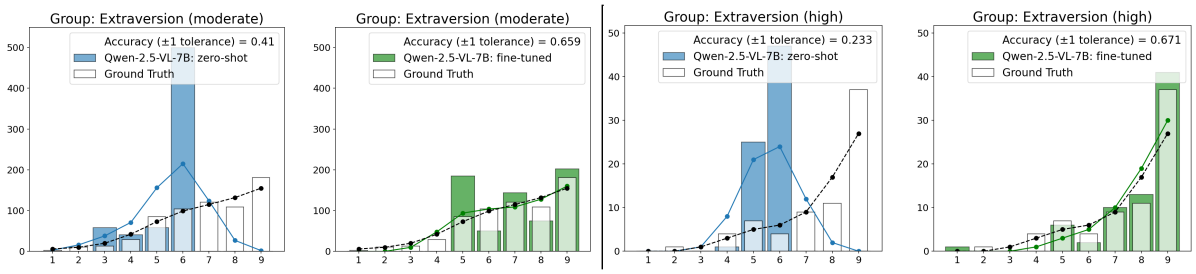


(b) Self-Efficacy Health Messages

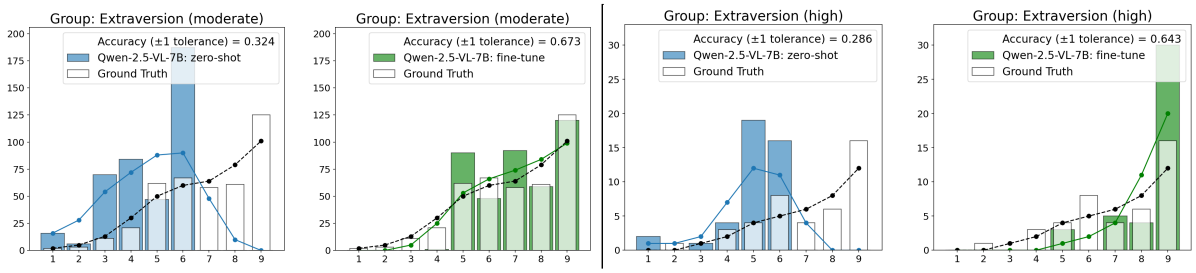


(c) Threatening/Fear-driven Health Messages

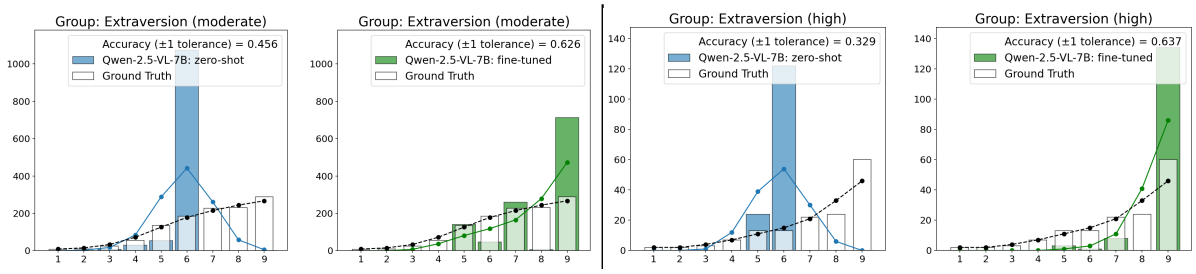
Figure 38: Sentimental response distributions of **Qwen2.5-VL** models to unseen communication strategies across **"Open-Mindedness" personality groups** (*moderate, high*). "*Low* Open-Mindedness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

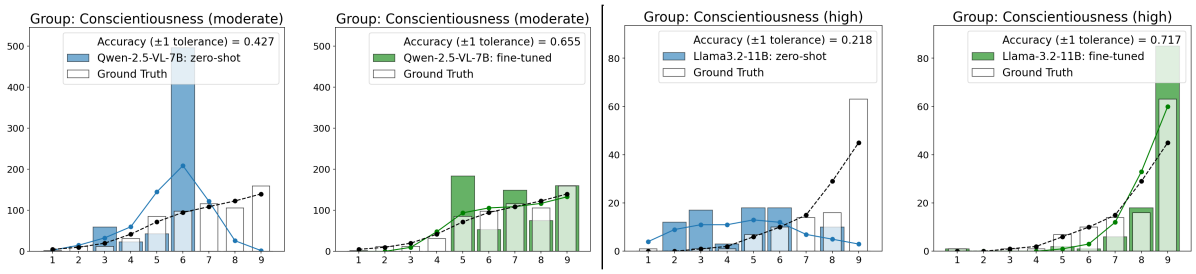


(b) Self-Efficacy Health Messages

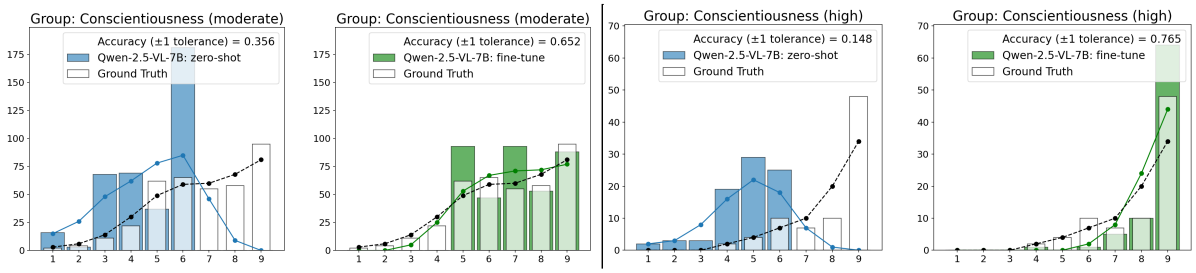


(c) Threatening/Fear-driven Health Messages

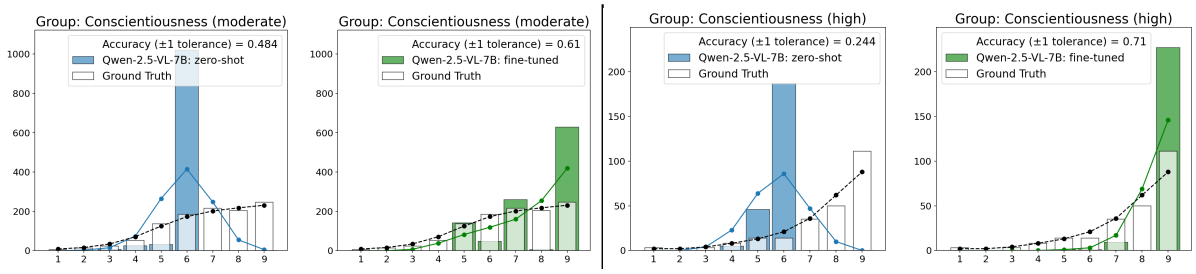
Figure 39: Sentimental response distributions of **Qwen2.5-VL** models to unseen communication strategies across **"Extraversion" personality groups** (*moderate, high*). "*Low* Extraversion" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

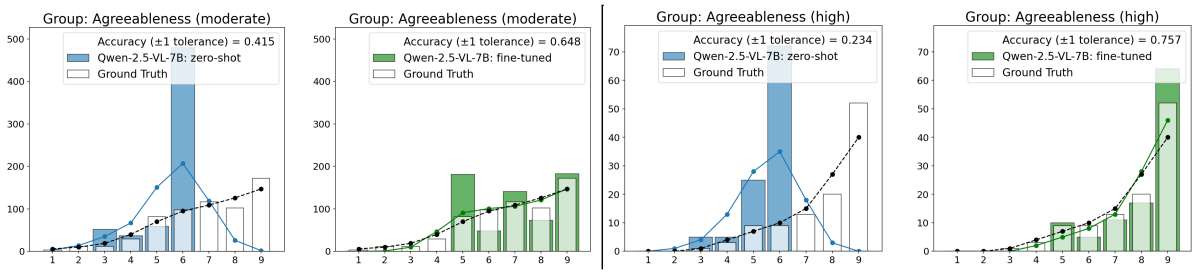


(b) Self-Efficacy Health Messages

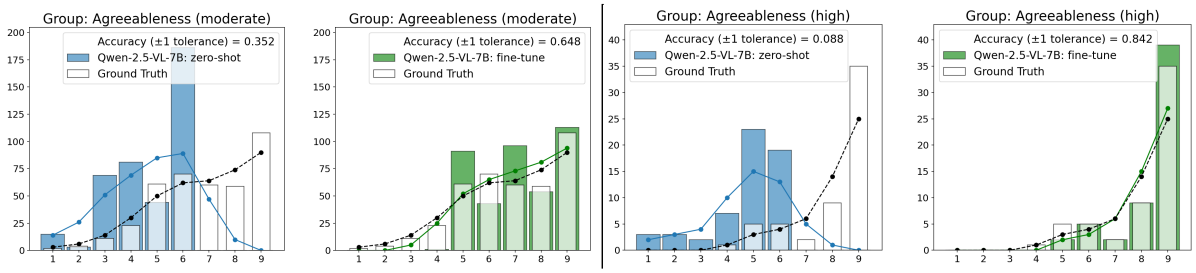


(c) Threatening/Fear-driven Health Messages

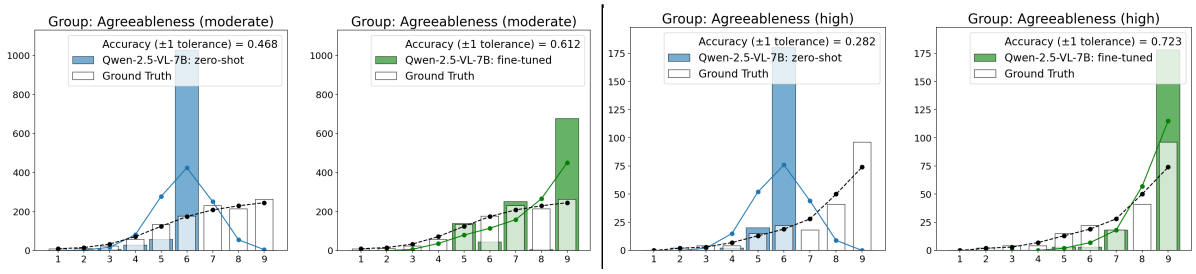
Figure 40: Sentimental response distributions of **Qwen2.5-VL** models to unseen communication strategies across **"Conscientiousness" personality groups** (*moderate, high*). "*Low* Conscientiousness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages

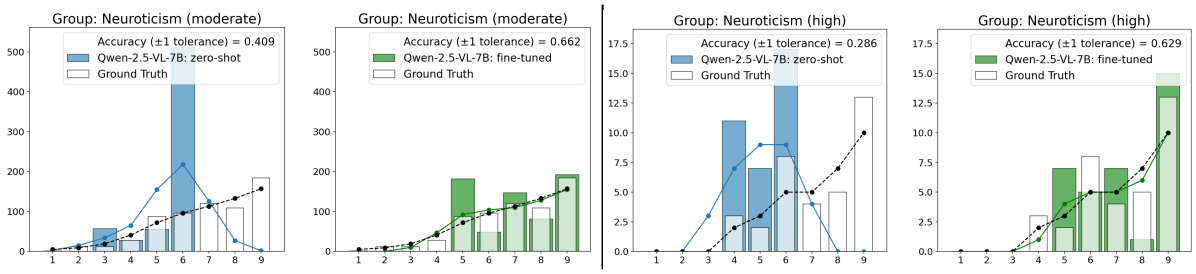


(b) Self-Efficacy Health Messages

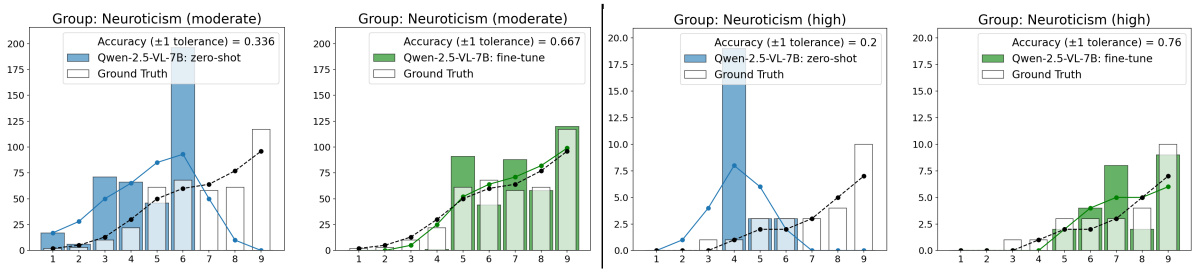


(c) Threatening/Fear-driven Health Messages

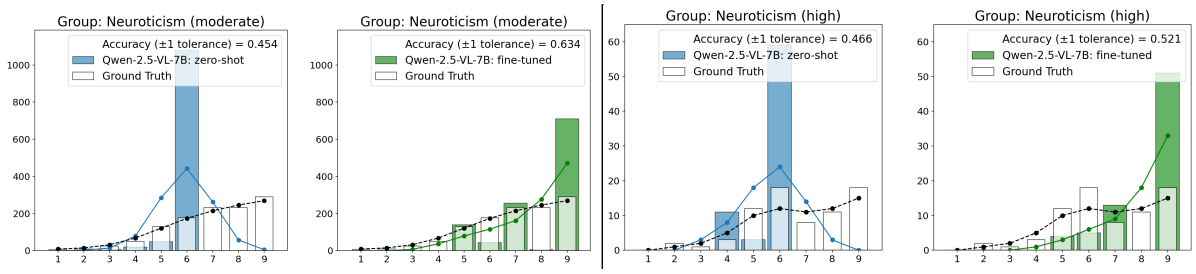
Figure 41: Sentimental response distributions of **Qwen2.5-VL** models to unseen communication strategies across **"Agreeableness" personality groups** (*moderate, high*). "*Low* Agreeableness" group is not included since it has less than 20 individuals in the test set. Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.



(a) Informational/Educational Health Messages



(b) Self-Efficacy Health Messages



(c) Threatening/Fear-driven Health Messages

Figure 42: Sentimental response distributions of **Qwen2.5-VL** models to unseen communication strategies across **"Neurocitism" personality groups** (*moderate, high*). Zero-shot (left) and trained (right) models are compared against the ground-truth distribution.

Ablation Study: Impact of Attribute Groups on Response Alignment under Unseen Communication Strategies

Model / Communication Strategy	Overall	Sentiment Prediction	Behavior Prediction	Emotion Prediction
"Informational/Neutral"	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2
Uniform Random Baseline	10.5 / 30.4 / 48.1	9.8 / 24.7 / 42.6	8.4 / 26.7 / 41.8	10.7 / 31.0 / 48.9
Gemma3-12B: zero-shot	13.7 / 43.8 / 68.4	16.1 / 48.2 / 84.6	13.2 / 43.8 / 74.4	11.1 / 37.6 / 61.6
Gemma3-12B: All Attributes	**32.7** / **57.6** / **75.0**	36.7 / **67.1** / 85.7	**38.1** / **65.5** / **84.9**	**29.1** / **51.9** / **69.5**
Gemma3-12B: No Demographic Attributes	29.5 / 55.1 / 73.3	33.9 / 65.0 / **86.0**	32.0 / 59.1 / 81.1	25.4 / 49.7 / 67.8
Gemma3-12B: No Psychometric Attributes	31.6 / 54.2 / 71.0	**37.7** / 63.8 / 79.5	33.0 / 56.0 / 77.0	27.5 / 50.2 / 66.7
"Self-Efficacy"	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2
Uniform Random Baseline	11.3 / 31.1 / 48.5	9.1 / 27.6 / 43.4	8.7 / 28.7 / 44.4	11.0 / 31.3 / 49.2
Gemma3-12B: zero-shot	13.9 / 45.9/ 71.0	16.3 / 45.9 / **87.7**	15.2 / 46.0 / 77.0	12.2 / 41.0 / 64.5
Gemma3-12B: All Attributes	**32.5** / **59.3** / **76.4**	35.2 / **67.8** / 84.7	32.6 / **64.6** / **82.9**	**30.0** / **55.2** / **72.2**
Gemma3-12B: No Demographic Attributes	31.7 / 56.9 / 75.3	37.0 / 65.7 / **85.4**	**36.2** / 62.7 / 81.6	28.1 / 52.6 / 71.0
Gemma3-12B: No Psychometric Attributes	31.4 / 56.5 / 73.3	**39.4** / 66.6 / 81.8	35.1 / 60.8 / 82.0	27.4 / 52.4 / 68.8
"Threatening/Fear-driven"	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2	exact / ±1 / ±2
Uniform Random Baseline	10.8 / 31.6 / 49.7	8.8 / 28.3 / 44.9	7.5 / 27.0 / 44.6	11.2 / 31.7 / 49.4
Gemma3-12B: zero-shot	14.5 / 43.0 / 66.7	16.6 / 49.0 / **86.8**	14.1 / 41.8 / 71.7	12.1 / 38.2 / 61.3
Gemma3-12B: All Attributes	**27.6** / **53.3** / 68.5	**34.3** / **66.9** / **83.9**	34.8 / **62.9** / 78.9	**23.4** / **47.7** / 62.9
Gemma3-12B: No Demographic Attributes	27.2 / 51.0 / **69.8**	33.3 / 62.2 / 82.7	33.4 / 56.5 / **79.2**	23.1 / 45.4 / **63.9**
Gemma3-12B: No Psychometric Attributes	27.4 / 48.1 / 63.3	33.7 / 62.7 / 80.6	**35.3** / 60.0 / 78.1	22.9 / 40.8 / 55.7

Table 3: Model's generalization performance (exact / ±1 / ±2 tolerance) on *unseen* messaging strategies under attribute ablations. The ±1 and ±2 tolerance metrics account for the imprecise nature of human sentimental and emotional responses, capturing distributional alignment with true responses without requiring exact matches. For each ablation setting, we retrain `Gemma-3:12B` with either demographic or psychometric attributes removed, isolating the contribution of each attribute group to generalization performance. Overall, **models trained with both demographic and psychometric attributes generally achieve the strongest performance**. Excluding either attribute group leads to consistent but moderate performance degradation rather than collapse, indicating that **LLMs can still align responses using partially available traits**. Despite these drops, ablation models substantially outperform random and zero-shot baselines, demonstrating robust response alignment even under incomplete attribute information. These gains over the baselines are statistically significant at $p \leq 0.0001$ under McNemar (McNemar, 1947) and cluster-robust pair boostrap (Efron and Tibshirani, 1994) ($B = 10{,}000$ and 95% CI) statistical tests.

H Ablation Studies on Attribute Group Contributions

Example of Demographic Attributes

```
Age: 35-44
Gender: Male
Assigned Sex at Birth: Male
Religious/Cultural Group: Christianity
Political Affiliation: Independent
Race/Ethnicity: White/Caucasian
Primary Language: English
First Language: English
Highest Education: Bachelor's degree
Employment Status: Employed full-time
Current Profession: Technology/Engineering
Annual Household Income: $100,000-$149,999
Marital Status: Married
Family Status: Married/Partnered with children
Zip Code: 92101
Disability: No
```

Example of Psychometric Attributes

```
Your personality test shows you have
(min score = 0; max score = 5):
Big-Five Trait Scores:
Extraversion: 2.8; Agreeableness: 3.0
Conscientiousness: 3.0; Neuroticism: 2.7
Open-Mindedness: 3.5

Facet Scores:
Sociability: 2.0; Assertiveness: 3.3
Compassion: 3.0; Respectfulness: 3.0
Trust: 3.0; Organization: 4.0
Productiveness: 3.3; Responsibility: 2.0
Anxiety: 2.0; Depression: 2.0
Emotional Volatility: 4.0
Intellectual Curiosity: 4.0
Aesthetic Sensitivity: 4.0
Creative Imagination: 4.0

You also have Locus of Control:
I'm my own boss: Applies mostly
If I work hard, I'll succeed: Applies somewhat
Others determine my actions: Applies somewhat
Fate impacts my plans: Applies somewhat
```

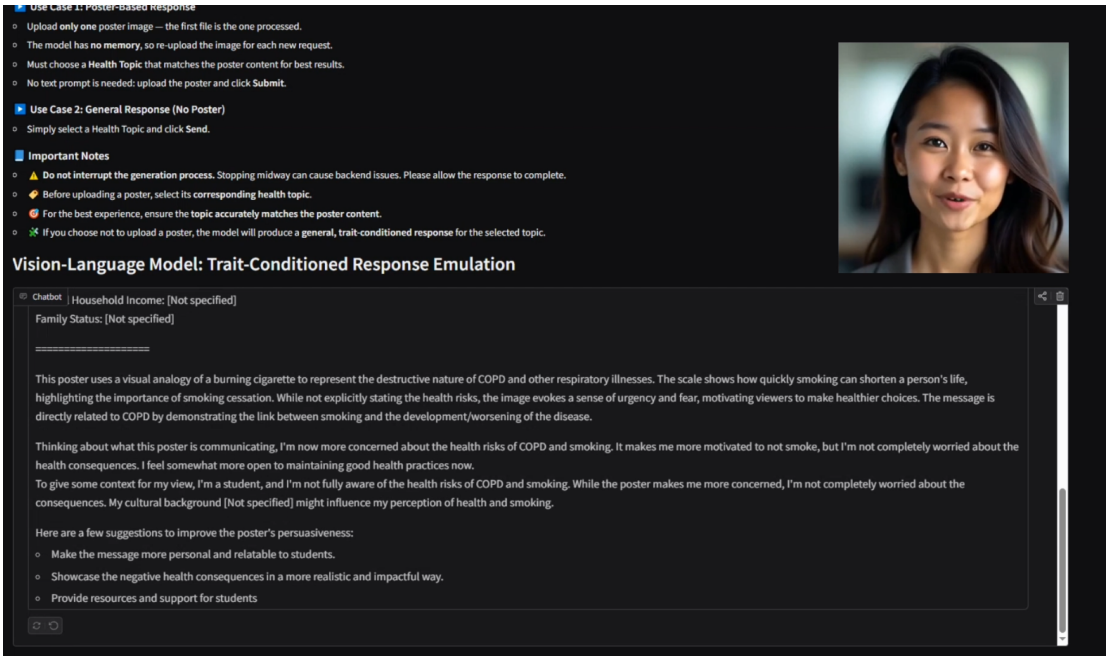
To quantify the contribution of demographic and psychometric attributes to response alignment performance, we conduct ablation studies in which each attribute group is removed in isolation. Specifically, we evaluate two ablation settings: **(1)** removal of demographic attributes and **(2)** removal of psychometric attributes. To ensure a fair and consistent comparison, attributes are excluded from both the alignment training and the inference stages. The generalization performance of the ablated models is reported in Table 3. We also show an example of demographic and psychometric attributes above.

Overall, we observe that when removing either demographic or psychometric attributes, the alignment training results in a consistent but moderate degradation in response prediction performance. These results indicate that both demographic and psychometric attribute groups contribute to maximizing the attribute-aware response prediction. Even under attribute ablation, aligned models generally outperform both random and zero-shot baselines across most metrics, highlighting the effectiveness of the alignment training. These findings demonstrate that both demographic and psychometric attributes contribute meaningfully to LLM-based response simulation, while alignment training remains effective even when one attribute group is not available.

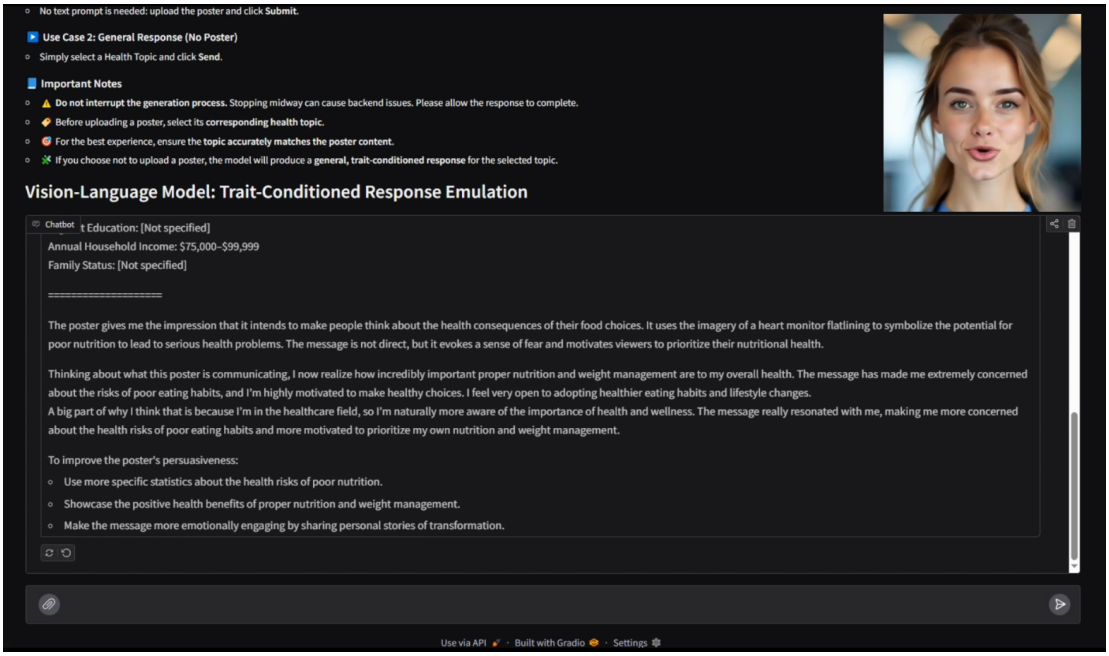
I Next Step: Interactive Avatar Integration

To further support user experience and student engagement, we are exploring the integration of real-time virtual avatar generation into CLAAS. While the current UI implementation primarily presents simulated community responses as text, an avatar-based interface can deliver the same trait-conditioned responses through spoken, conversational feedback. This extension follows a simple pipeline: the generated textual response is converted into a short script, synthesized into speech, and rendered through a talking-avatar interface whose appearance is conditioned on user-specified demographic traits. An example UI is shown in Figure 43. *We have included the demo recordings of the avatar UI in the Supplementary Materials along with the codebase*.

This feature is motivated by student interest in animated-avatar feedback across the Fall 2025 and Spring 2026 pilots, where 50% and 75% of respondents, respectively, indicated that such a feature would be helpful, as shown in Figure 44. The interactive avatar demonstrates the extensibility of CLAAS and illustrates how attribute-aware response simulation can be delivered through richer modalities in future deployments. Such interfaces may help students engage more naturally with simulated community perspectives and better reflect on how different target populations may receive their campaign messages. Future work will evaluate whether avatar-based interaction improves engagement, reflection, and learning outcomes while maintaining transparency and appropriate trust calibration.

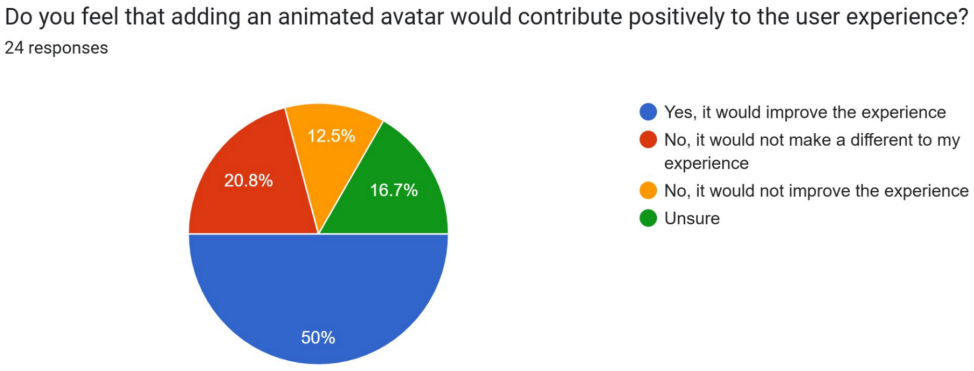


(a) Targeted profile: Female, 18–24 years old, Asian, Occupation: Student.

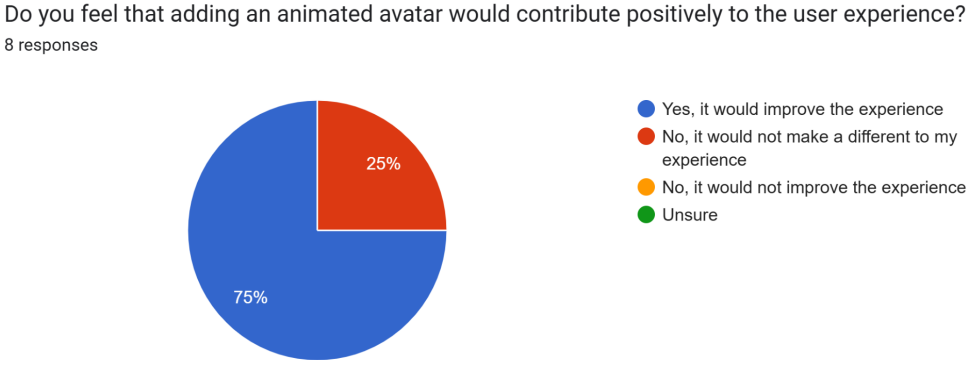


(b) Targeted profile: Female, 18–24 years old, Caucasian, Occupation: Healthcare/Medical.

Figure 43: **Example interactive avatar UI.** The avatar is dynamically generated based on user-specified target demographic traits. Full video demos are included in the ***Supplementary Materials***.



(a) Students' preference for an animated avatar feature, Fall 2025 pilot.



(b) Students' preference for an animated avatar feature, Spring 2026 pilot.

Figure 44: **Student preference for the interactive avatar feature.** Pilot feedback suggests majority of students show interest in receiving simulated responses through an animated avatar interface.