Scientific Journal of King Faisal University:
Humanities and Management Sciences

GenAI Agent for Automated Analysis and Personalization of Drug Prevention Campaigns

Mohammed Aljaafari and Shaymaa E. Sorour
Department of Management Information Systems, School of Business, King Faisal University, Al-Ahsa, Saudi Arabia

LINK	RECEIVED	ACCEPTED	PUBLISHED ONLINE	ASSIGNED TO AN ISSUE
https://doi.org/10.37575/h/mng/250075	13/12/2025	03/02/2026	03/02/2026	01/03/2026
NO. OF WORDS	**NO. OF PAGES**	**YEAR**	**VOLUME**	**ISSUE**
8110	10	2026	27	1

ABSTRACT

This study introduces a generative artificial intelligence (GenAI) agent designed to autonomously evaluate, optimize, and personalize drug prevention campaigns across Facebook, Reddit, Instagram, and Twitter (X) using a 45,000-post multi-platform awareness corpus. Five state-of-the-art large language models, GPT-5-mini, Claude 3.5, Gemini 2.0, Qwen 2.5, and Mixtral, were examined under six structured prompting families, including Original, Role-based, Two-Stage Explicit Sensemaking, and Case-Based Adaptive Reasoning in short and long variants. Model outputs were assessed using a tri-metric framework comprising the Educational Rate (Edu-R), Violation Rate (Vio-R), and Misleading Awareness Score (MAS), supported by classical discrimination and agreement measures, including ROC–AUC and Cohen's Kappa, as well as influence-spread simulation. Results demonstrate that GPT-5-mini exhibits the strongest overall performance, achieving 95.10% accuracy, 96.22% precision, 94.55% recall, and 95.44% F1 score. Structured prompting substantially improved alignment and safety across all models, increasing GPT-5-mini's Edu-R from 78.12% under minimal instructions to over 95% under agent-based prompting. The Vio-Rs were reduced to low single-digit values, corresponding to approximately 96%–99% safety-aligned outputs. Influence-spread simulations further showed that cognitively rich prompts significantly enhance message diffusion, particularly in demographic clusters. The proposed GenAI agent establishes a scalable, evidence-driven foundation for real-time evaluation and personalization of drug prevention campaigns.

KEYWORDS
Drug-awareness, educational rate, generative AI, misleading awareness, social media, violation rate

CITATION
Aljaafari. M. and Sorour. S.E. (2026). GenAI agent for automated analysis and personalization of drug prevention campaigns. *Scientific Journal of King Faisal University: Humanities and Management Sciences*, **27**(1), 68–77. DOI: 10.37575/h/mng/250075

1. Introduction

The global escalation in drug abuse, particularly among adolescents and young adults, represents an urgent public health crisis that demands innovative, scalable, and evidence-based prevention strategies (Olawade *et al.*, 2023; Panteli *et al.*, 2025). Traditional awareness campaigns relying on lectures and printed materials often lack the adaptability, reach, and real-time feedback mechanisms necessary to engage digitally native youth and respond swiftly to evolving drug-use trends (Khakpaki and Sepehri, 2025; Nishan, 2025). The expanding ubiquity of social media platforms such as Facebook, Reddit, and Twitter (X) offers unprecedented channels for health communication; however, leveraging these mediums effectively requires advanced analytic and personalization capabilities that transcend conventional approaches (Li *et al.*, 2025; Khosravi *et al.*, 2024).

Recent advances in generative artificial intelligence (GenAI), driven by next-generation large language models (LLMs), including GPT-5-mini, Claude 3.5, Gemini 2.0, Qwen 2.5, and Mixtral, provide transformative opportunities for the design, analysis, and optimization of public health campaigns (Panteli *et al.*, 2025). These models exhibit enhanced reasoning, multimodal understanding, and contextual awareness, which enable near real-time processing and interpretation of vast textual and multimedia content generated on digital platforms (Bandeira *et al*, 2025). By harnessing powerful natural language processing (NLP) and sentiment analysis techniques, GenAI agents can autonomously assess the effectiveness of drug-prevention messages, quantify public engagement, and adapt campaign content dynamically to demographic, cultural, and behavioral patterns (Villanueva *et al.*, 2025).

The central innovation of this study lies in the integration of state-of-the-art GenAI models to systematically compare digital social media campaigns against traditional approaches, enabling continuous measurement of relevance, sentiment, and behavioral impact. This approach incorporates the influential role of youth and community leaders in message diffusion, capitalizing on trust and peer influence to enhance campaign credibility and uptake (Nwanakwaugwu *et al.*, 2025; Plackett *et al.*, 2025). Simultaneously, the technology autonomously generates comprehensive qualitative and quantitative reports, supporting policymakers and health organizations with a robust, data-driven foundation for evidence-based decision-making and resource optimization.

Significant prior work underscores the efficacy of AI in early detection of substance misuse risks through examination of social media data, personal health records, and behavioral patterns, enabling personalized intervention planning (Deng *et al.*, 2024). Sentiment and topic modeling have proven particularly valuable in unravelling public attitudes towards drug use, elucidating factors that influence perceptions and stigma surrounding interventions (Bandeira *et al.*, 2025; Villanueva *et al.*, 2025). However, despite these advances, current systems often lack real-time adaptivity and personalized message generation capabilities, which are essential for effectively engaging diverse populations.

The proposed GenAI framework addresses these critical gaps by leveraging the next wave of GenAI architectures to not only evaluate but also personalize prevention campaigns in an ongoing, scalable fashion. By contextualizing messages according to socio-demographic variables and cultural nuances, the system advances the precision and impact of awareness initiatives beyond static and generic content delivery. This technological integration, reinforced by the amplification effects of trusted influencers, holds significant promise for reducing drug-related harms among youth populations.

This paper introduces a novel approach to addressing the challenges of AI-driven drug-awareness evaluation. The approach has several unique contributions, which are highlighted below.

- A 45,000-post multi-platform corpus spanning Facebook, Reddit, Instagram, and Twitter (X) is developed to capture heterogeneous awareness content, risk cues, and misinformation patterns reflective of

Corresponding Author: Mohammed Aljaafari 009660557122210, mahaljaafari@kfu.edu.sa

real-world social-media environments.
- A structured GenAI Agent, incorporating five advanced prompting families, is designed to enhance reasoning stability, safety alignment, and contextual fidelity across GPT-5-mini, Claude 3.5, Gemini 2.0, Qwen 2.5, and Mixtral.
- A three-dimensional evaluation framework comprising Educational Rate (Edu-R), Violation Rate (Vio-R), and the Misleading Awareness Score (MAS) is introduced to provide a rigorous and quantifiable assessment of educational quality, safety compliance, and misinformation risk.
- A segmentation-driven evaluation demonstrates the system's capacity to generate context-sensitive responses for distinct user profiles, particularly differentiating general awareness inquiries from early addiction–risk cases.
- Influence-spread modeling based on diffusion simulation is employed to estimate real-world propagation dynamics, demonstrating that the proposed agent-based prompting strategies substantially enhance the dissemination and potential impact of awareness messages.

2. Related Work

Recent years have witnessed a significant expansion of AI and LLMs within public health, particularly for drug abuse prevention and awareness campaigns. Early applications of AI, including machine learning–driven text mining and explainable models, have supported epidemiological surveillance, targeted intervention, and automated sentiment analysis on various digital platforms (Panteli *et al.*, 2025; Ye *et al.*, 2025; Zhao *et al.*, 2021). GenAI architectures such as GPT-series and multimodal deep learning networks now enable real-time scrutiny of campaign effectiveness, participant engagement, and cultural resonance through advanced NLP pipelines (Ahmad *et al.*, 2025; Li *et al.*, 2025; Maharjan *et al.*, 2025; Villanueva *et al.*, 2025).

Systematic reviews confirm that deep learning approaches for topic and sentiment modeling on social media outperform traditional keyword-based analytics, yielding more accurate representations of youth attitudes towards substance use (Villanueva *et al.*, 2025). Community health workers and influencers, especially when empowered by AI assistant technologies, have been shown to amplify the credibility and reach of digital campaigns, thus broadening their behavioral impact among vulnerable groups (Bharel *et al.*, 2024; Khakpaki and Sepehri, 2025).

Responsible deployment of AI in public health requires robust attention to ethics, bias reduction, and model transparency. Frameworks integrating explainable machine learning and multimodal data mining facilitate policymakers' ability to identify gaps and continuously optimize strategies (Panteli *et al.*, 2025; Uddin *et al.*, 2025; Ye *et al.*, 2025). The literature also highlights the importance of tailoring prevention messages to demographic and cultural nuances using generative models, which have demonstrated early success in improving campaign adaptivity and outcomes.

Overall, mounting evidence supports the utility of GenAI agents as scalable tools for automating, evaluating, and personalizing drug prevention campaigns — maximizing relevance, engagement, and efficacy within diverse populations (Panteli *et al.*, 2025; Li *et al.*, 2025; Villanueva *et al.*, 2025).

Although AI and LLMs have been applied to substance-use surveillance, prior studies are constrained by limited datasets, single-task evaluations, and insufficient attention to safety or audience-specific behaviors. This study overcomes these limitations through a unified evaluation framework that integrates a multi-platform corpus, a structured GenAI agent with five prompting strategies, and a tri-metric assessment model (Edu-R, Vio-R, MAS) for jointly measuring educational quality, safety alignment, and misinformation risk. The framework further incorporates segmentation-responsive analysis to capture differential system behavior across high-impact user groups, including individuals showing early addiction signals. Finally, influence-spread diffusion modeling provides an analytically grounded estimate of real-world message propagation. Together, these elements establish a concise, safety-aware, and audience-adaptive paradigm not previously achieved in drug-prevention research.

3. Data Sources and Collection Overview

This study employs a large, multi-platform dataset comprising 45,000 publicly available social media posts collected from Facebook, Twitter (X), Reddit, and Instagram. Table 1 summarizes the number of posts from each platform and the corresponding content styles. These posts capture online discourse related to drug prevention, public health awareness, and community engagement. The construction of such a heterogeneous corpus ensures that the proposed GenAI agent is trained and evaluated on authentic, demographically diverse, and linguistically rich digital communications. This approach is consistent with recommendations from contemporary review studies that emphasize the importance of real-world, multi-source datasets for developing robust natural language-understanding systems (Alamoodi *et al.*, 2020; Nasser and Abu-Naser 2024; Zhu *et al.*, 2022).

The dataset covers the full landscape of online awareness messaging, including public opinions, emotional responses, institutional campaign announcements, community dialogues, harm-reduction guidance, and youth-oriented discussions related to substance use and national prevention initiatives.

The corpus includes posts related to:

- Drug-prevention messages and public-health communication.
- Governmental and non-governmental organisation (NGO) anti-drug campaigns.
- Community and youth discussions about substance use.
- Reactions to national policies and educational initiatives.
- Supportive, motivational, and engagement-driven messages.
- Harm-reduction advice, safety information, and awareness materials.

For experimental control, the corpus was partitioned at the post level into disjoint training, validation, and test sets. A total of 36,000 posts were used for model conditioning and prompt development, while 9,000 posts were reserved as an independent, held-out test set. Data partitioning was performed prior to any preprocessing or model interaction to prevent cross-set contamination. A validation subset (10% of the training data) was used exclusively for prompt and configuration tuning. All reported results correspond to evaluation on the held-out test set.

The following subsections provide a detailed breakdown of each data platform.

Table 1. Dataset distributions by platform

Platform	Posts	Contribution	Content Style
Twitter (X)	22,000	48.9%	Short, emotional, fast-paced
Reddit	15,500	34.4%	Long, detailed, discussion-based
Instagram	3,500	7.8%	Emojis, captions, mixed-language
Facebook	4,000	8.9%	Formal and community campaign posts
Total	45,000	100%	Multi-platform, diverse public discourse

3.1. Twitter (X):

Twitter (X) constitutes the largest component of the dataset. The platform's high velocity, public accessibility, and widespread use of hashtags make it an essential source of real-time awareness and opinion dynamics. Posts were collected using keyword- and hashtag-based retrieval strategies specifically targeting prevention-oriented discourse (AlBarrak and Sorour, 2024; Olivares-De la Fuente *et al.*, 2025).

Linguistic and Interactional Features:

- Short-form messages (≤280 characters)
- High density of hashtags and trending cues
- Emotionally expressive and reactive content
- Engagement indicators (retweets, replies, quote-tweets)

- Suitable for fine-grained sentiment and stance analysis

Twitter's (X's) rapid, conversational style provides an essential layer of temporal and emotional variability that is crucial for training generative agents on real-world public reactions.

3.2. Reddit:

Reddit represents the second-largest data source, offering long-form, community-driven discussions across topic-specific subreddits. Its structured conversational threads allow for deep semantic exploration and nuanced thematic mapping (Lendvai, 2025; Sorour and Almusallam, 2025).

Typical Subreddits:

- r/drugs
- r/healthpromotion
- r/addiction
- r/SaudiArabia
- r/medicine

Key Linguistic and Interactional Features:

- Paragraph-length narratives
- High semantic richness and contextual depth
- Personal stories, peer guidance, debates
- Rich source for topic modeling and emotion variability

Reddit's contribution is critical for capturing complex reasoning, debates, and long-form reflections often absent in microtext platforms.

3.3. Instagram:

Instagram posts were collected from publicly accessible captions, comments, and awareness campaign hashtags. This platform provides content enriched with creative language use, multilingual expression, and emotive markers (Bharel *et al.*, 2024; Brandao *et al.*, 2024).

Linguistic and Interactional Features:

- Medium-length captions with narrative style
- Frequent use of emojis and stylistic elements
- Arabic–English bilingual and dialectal content
- Influencer-driven engagement

Instagram adds emotional tone and cultural nuance, enhancing the linguistic diversity required for robust language modeling.

3.4. Facebook:

Facebook contributes a smaller but highly structured subset of posts. These entries primarily originate from official public health pages, universities, NGOs, and community programs (Ho *et al.*, 2025).

Linguistic and Interactional Features:

- Formal campaign announcements
- Public comments facilitating two-way communication
- Lower posting frequency but higher content clarity
- Valuable for identifying institutional messaging strategies

Facebook enhances the dataset with clear, policy-aligned, professionally curated prevention content that complements informal public discourse.

4. Methodology

This study employs a structured methodological approach combining multi-platform data collection, standardized text preprocessing, and advanced prompting strategies to evaluate the performance of five state-of-the-art GenAI models in drug-awareness analysis. Model outputs are assessed using a three-dimensional evaluation framework (Edu-R, Vio-R, and MAS), supported by segmentation and influence-spread analysis to examine audience-specific responses and real-world diffusion potential.

The complete corpus of 45,000 social media posts was partitioned at the post level using a fixed 80:20 split, resulting in 36,000 posts for training and 9,000 posts for testing. Data splitting was performed prior to any preprocessing, prompt construction, augmentation, or model interaction to prevent information leakage. A validation subset comprising 10% of the training data was reserved exclusively for prompt and configuration selection.

4.1. Model Selection and Specification:

The study evaluates five proposed LLMs, (as summarized in Table 2), all accessed through their official application programming interfaces. Model identifiers, release states, and access periods were fixed throughout the experimental phase and are explicitly reported to ensure reproducibility and temporal consistency. No model updates or endpoint changes were applied during evaluation.

Table 2. Specification of the five proposed LLMs

Model	Provider	Model Identifier	Snapshot / Release	Access Period
GPT-5-mini[1]	OpenAI	GPT-5-mini	Stable snapshot (2025)	Jun–Dec, 2025
Claude 3.5[2]	Anthropic	claude-3.5	Stable release	
Gemini 2.0[3]	Google DeepMind	gemini-2.0-pro	Stable release	
Qwen 2.5[4]	Alibaba Cloud / HF	qwen-2.5	2024–2025 release	
Mixtral[5]	Mistral AI	mixtral-8×7b	Public release	

4.2. Data Acquisition and Integration:

In the current study, social media posts related to drug-awareness themes were collected from four major platforms (Twitter [X], Reddit, Instagram, and Facebook). Each platform dataset is represented as D_i, and a unified dataset was created via horizontal concatenation:

$$D_{\text{merged}} = \bigcup_{i=1}^{4} D_i, \tag{1}$$

where D_i is dataset from platform *I*, D_{merged} is final merged dataset for all experiments, and $i = 1 \ldots 5$ is index of the data source.

4.3. Text Preprocessing:

Text preprocessing followed standard NLP pipelines (Eisenstein, 2019; Manning and Schutze, 1999). Raw text (x) was normalised using a composite cleaning function:

$$x' = f_{\text{clean}}(x) = f_{\text{lower}}\left(f_{\text{stop}}\left(f_{\text{url}}\left(f_{\text{emoji}}(x)\right)\right)\right), \tag{2}$$

where x is raw input text, x' is cleaned text, f_{clean} is full cleaning pipeline, f_{emoji} removes emojis, f_{url} removes URLs, f_{stop} removes stop words, and f_{lower} converts text to lowercase.

Tokenization was conducted as follows:

$$t = \text{Tokenizer}(x'), \tag{3}$$

where t is token sequence and Tokenizer $(\cdot)$ is model-compatible tokenizer.

4.4. Dataset Annotation and Labeling Protocol:

To ensure the reliability, validity, and reproducibility of the ground-truth labels used for model evaluation, a structured manual annotation protocol was adopted for the social media corpus. The dataset was annotated to classify awareness-related posts into three ordinal categories: High-Value Awareness, Moderately Valuable Awareness, and Low-Value Awareness, following established principles of systematic content analysis (Krippendorff, 2018).

1 https://platform.openai.com/docs/models
2 https://docs.anthropic.com
3 https://ai.google.dev
4 https://huggingface.co/Qwen
5 https://docs.mistral.ai

4.4.1. Annotator Composition and Expertise

The annotation process was conducted by five independent annotators with prior experience in public health communication, health-related content analysis, and social media moderation. All annotators possessed domain familiarity with substance-use prevention discourse and were proficient in the linguistic styles represented in the dataset, including informal and platform-specific expressions. The use of multiple trained annotators is a recommended practice to mitigate individual bias and improve labeling robustness (Krippendorff, 2018).

4.4.2. Annotation Guidelines

Annotators were provided with a detailed guideline document specifying category definitions, inclusion and exclusion criteria, and representative examples.

- High-Value Awareness posts were defined as content that is accurate, informative, prevention oriented, and promotes constructive understanding or protective behavior.
- Moderately Valuable Awareness posts conveyed partial or generic information without explicit preventive guidance or actionable recommendations.
- Low-Value Awareness posts included vague, misleading, sensationalized, or weakly informative content, as well as posts that lacked clear educational intent.

Explicit category definitions and edge-case clarifications were used to reduce subjective interpretation and enhance annotation consistency, in line with best practices in content analysis methodology (Krippendorff, 2018).

4.4.3. Labeling Procedure

Each post was independently annotated by at least two annotators, ensuring redundancy in label assignment. Annotation tasks were balanced across annotators to minimize fatigue effects. Annotators were blinded to all model outputs, preserving the independence of the ground-truth labels and preventing confirmation bias.

4.4.4. Inter-Rater Agreement Metrics

Inter-rater reliability was quantified using Cohen's Kappa (κ), a chance-corrected agreement measure widely adopted for nominal-scale annotations (Cohen, 1960). Cohen's κ is defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{4}$$

where p_o denotes the observed agreement between annotators and p_e represents the expected agreement by chance.

For annotation tasks involving multiple annotator pairs, the overall agreement was summarized using the average pairwise κ, computed as follows:

$$\kappa_{\text{avg}} = \frac{1}{M}\sum_{i=1}^{M} \kappa_i, \tag{5}$$

where κ_i denotes the Cohen's κ value for annotator pair i, and M is the total number of annotator pairs. This formulation provides a statistically grounded estimate of overall annotation consistency (Cohen, 1960).

4.4.5. Disagreement Resolution

In cases of label disagreement, a consensus-based adjudication process was applied. Conflicting annotations were jointly reviewed by the involved annotators with reference to the predefined guidelines. If consensus could not be reached, the final label was determined by a senior annotator with domain expertise. This adjudication strategy ensured conceptual consistency while maintaining methodological rigor, as recommended in systematic content-analysis frameworks (Krippendorff, 2018).

4.5. Language Model Configuration:

Each LLM (GPT-5-mini, Claude 3.5, Gemini 2.0, Qwen 2.5, Mixtral) was modeled as a conditional generative probability distribution (Brown *et al.*, 2020; Vaswani *et al.*, 2017):

$$M_\theta(y \mid x) = \prod_{t=1}^{T} \quad P_\theta(y_t \mid y_{<t}, x), \tag{6}$$

where M_θ denotes model with parameters θ, y denotes output token sequence, x denotes input text, y_t is token at time t, $y_{<t}$ denotes all tokens before t, T is output length, and $P_\theta(\cdot)$ denotes model probability distribution.

Temperature-based sampling was applied:

$$P_{\theta,T}(y_t) = \frac{P_\theta(y_t)^{\frac{1}{T}}}{\sum_k \quad P_\theta(y_k)^{\frac{1}{T}}}, \tag{7}$$

where T is sampling temperature and k is Vocabulary index.

4.6. Prompt-Length and Reasoning-Depth Variants in LLM Alignment Evaluation:

To assess how prompt length and reasoning structure affect model behavior, each prompting family was implemented in short (S) and long (L) variants: ORIG_S/ORIG_L, 2SE_S/2SE_L, and CA2NE_S/CA2NE_L. The ORIG variants provide minimal versus extended instructions, reflecting evidence that prompt verbosity influences LLM stability and interpretive accuracy (Brown *et al.*, 2020; Villanueva *et al.*, 2025). The Two-Stage Explicit Sensemaking (2SE) variants apply concise or expanded explicit sensemaking steps that strengthen reasoning coherence and reduce unsafe outputs (Kojima *et al.*, 2022; Hendrycks et al., 2020). The Case-Based Adaptive Reasoning (CA2NE) variants use short or detailed case-based cues to enhance contextual grounding in public health communication (Bharel *et al.*, 2024; Solaiman and Dennison, 2021). These six configurations collectively enable a compact examination of how linguistic depth and cognitive scaffolding shape educational alignment and safety performance across models.

4.6.1. Short and Long Prompt Variants

To investigate how instructional verbosity influences model alignment and reasoning stability, each prompting family was implemented in two structured variants: a short form and a long form. The short variants ORG_S and ROLE_S contain only the essential directive or role assignment, offering minimal contextual framing. Prior studies have shown that such low-information prompting increases interpretive variability and reduces output stability in LLMs, particularly when domains require high sensitivity or safety-aware reasoning (Brown *et al.*, 2020; Kojima *et al.*, 2022; Villanueva *et al.*, 2025). Therefore, these short variants function as baseline conditions to assess intrinsic model behavior under minimal instructional guidance.

In contrast, the long variants, ORG_L and ROLE_L, expand the instructional content through richer semantic framing, explicit contextual cues, and more detailed expert-role descriptions while deliberately avoiding explicit reasoning chains. Empirical evidence demonstrates that enhanced prompt elaboration improves LLM robustness, strengthens safety adherence, and reduces the risk of harmful or misleading outputs in public health communication contexts (Bharel *et al.*, 2024; Hendrycks *et al.*, 2020; Solaiman and Dennison, 2021). By comparing short and long variants under non-agent conditions, this study offers a controlled assessment of how prompt length alone modulates educational accuracy, safety alignment, and interpretive precision in drug-awareness message generation.

4.6.2. Prompt-Length and Reasoning-Depth Variants

To systematically evaluate how linguistic depth and reasoning structure influence model performance, each prompting family was

operationalized in short (S) and long (L) variants, yielding six structured configurations. The ORIG_S and ORIG_L prompts represent minimal and extended forms of the Original Prompt, respectively; ORIG_S provides only the essential task instruction, whereas ORIG_L introduces a more elaborated formulation without adding explicit reasoning. Prior work shows that prompt length alone can meaningfully modulate model interpretability, stability, and adherence to task constraints (Brown *et al.*, 2020; Miao *et al.*, 2024). These two variants, therefore, serve as a baseline for examining how instructional verbosity influences drug-awareness message generation.

The 2SE_S and 2SE_L variants correspond to the short and long forms of Two-Stage Explicit Sensemaking, a prompting framework that guides the model through structured reasoning steps such as reflection, justification, and decision synthesis; 2SE_S offers a concise scaffold, whereas 2SE_L expands these steps into a more detailed cognitive pathway. Research in structured prompting indicates that explicit reasoning supports greater coherence, reduces harmful outputs, and improves policy-aligned decision-making in LLMs (Hendrycks *et al.*, 2020; Kojima *et al.*, 2022). These variants allow the study to quantify how expanded reasoning depth affects educational and safety outcomes.

Finally, CA2NE_S and CA2NE_L extend the Case-Based Sensemaking framework; CA2NE_S provides a compact scenario-based reasoning cue, whereas CA2NE_L enriches the scenario with additional contextual, behavioral, and situational details. Case-based prompting has been shown to improve contextual fidelity and domain sensitivity, particularly in health and safety communication where situational grounding enhances model reliability (Bharel *et al.*, 2024; Solaiman and Dennison, 2021). Through these two variants, the study isolates the effect of expanding case complexity on safety alignment and awareness quality, offering a rigorous comparison of prompt-length and reasoning-depth interactions across all evaluated models.

4.7. Awareness Metrics (Edu-R, Vio-R, MAS):

Three complementary metrics were used to quantify educational accuracy, safety alignment, and misinformation risk.

- The Edu-R measures the proportion of accurate, prevention-aligned outputs:

$$\text{Edu-R} = \frac{N_{\text{edu}}}{N_{\text{total}}} \times 100, \tag{8}$$

- The Vio-R captures the frequency of unsafe or misleading responses:

$$\text{Vio-R} = \frac{N_{\text{vio}}}{N_{\text{total}}} \times 100, \tag{9}$$

- The MAS) normalizes unsafe behavior relative to educational performance:

$$\text{MAS} = \frac{\text{Vio-R}}{\text{Edu-R}+\varepsilon}. \tag{10}$$

These metrics are grounded in safety-evaluation methodologies widely adopted in alignment research, particularly for high-risk public health communication contexts (Hendrycks *et al.*, 2020; Solaiman and Dennison, 2021).

4.8. Model Discrimination Metrics:

The discrimination metrics presented in this section complement the primary awareness-quality indicators (Edu-R, Vio-R, MAS) by evaluating how effectively each model distinguishes educationally appropriate outputs from harmful or misleading ones across varying decision thresholds. Sensitivity and specificity describe the model's ability to correctly identify aligned content while rejecting unsafe responses, and the receiver operating characteristic–area under the curve (ROC–AUC) measure provides a threshold-independent view of overall discriminative reliability. When combined with the tri-metric awareness framework, these classical detection metrics offer a unified methodological structure that captures content-level alignment and the underlying decision-separation behavior of the evaluated models. Such integrated evaluation frameworks are widely recommended in AI-safety and digital-health research to ensure dependable model behavior in high-stakes public health communication settings (Amann *et al.*, 2020; Hendrycks *et al.*, 2020; Panteli *et al.*, 2025; Solaiman and Dennison, 2021; Ye *et al.*, 2025).

- Sensitivity (Recall / True Positive Rate): Sensitivity evaluates the model's capacity to correctly identify educationally valid outputs:

$$\text{TPR}(\tau) = \frac{TP(\tau)}{TP(\tau)+FN(\tau)}. \tag{11}$$

- Specificity (True Negative Rate): Specificity measures the ability to identify harmful or misleading outputs:

$$\text{TNR}(\tau) = \frac{TN(\tau)}{TN(\tau)+FP(\tau)}. \tag{12}$$

- False Positive Rate (FPR): The False Positive Rate quantifies erroneous acceptance of unsafe content:

$$\text{FPR}(\tau) = \frac{FP(\tau)}{FP(\tau)+TN(\tau)} \tag{13}$$

- Area Under the ROC Curve: Discrimination performance across thresholds is summarized as follows:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}). \tag{14}$$

Values approaching 1.0 indicate strong differentiation between educational and misleading content, consistent with findings in LLM reliability research (Kojima *et al.*, 2022; Miao *et al.*, 2024).

4.8.1. Grouping Strategy for Responsiveness Analysis

To examine heterogeneity in responsiveness to prompting strategies, the dataset was partitioned into three analytically distinct groups based on content structure, interaction depth, and network embedding, rather than demographic characteristics, to preserve user privacy and comply with ethical data-use constraints.

- Group 1 (Low-Context / Broadcast-Oriented Content): This group comprises short, declarative posts with minimal contextual dependency and limited conversational depth. Content in this group is typically one-directional, delivering awareness or informational messages with weak engagement signals and negligible interaction history.
- Group 2 (Moderate-Context / Discussion-Oriented Content): This group includes posts embedded within short discussion threads, exhibiting moderate contextual dependency and partial interaction continuity. Posts often contain emotional, argumentative, or explanatory elements, with observable but limited engagement dynamics.
- Group 3 (High-Context / Network-Embedded Content): This group consists of posts situated within dense conversational or diffusion networks, characterized by multi-turn interactions, rich contextual cues, and elevated exposure to social influence mechanisms. Content in this group frequently participates in cascaded sharing, reply chains, or debate-intensive environments, resulting in greater sensitivity to prompting strategies and higher diffusion potential.

4.9. Influence-Spread Modeling:

Influence-spread analysis was conducted using the independent cascade (IC) model to estimate how prompting strategies affect the dissemination of awareness messages. This framework reflects well-established diffusion principles wherein activation spreads through a network based on local influence probabilities (Bharel *et al.*, 2024).

4.9.1. Network Construction for Modeling

To support influence-spread analysis under the IC framework, a data-

driven synthetic social network was constructed to approximate realistic online interaction patterns observed in social media environments. Within this network, nodes correspond to unique user accounts, whereas directed edges represent interaction relationships, including replies, mentions, and reposts, derived from the underlying multi-platform dataset.

The constructed network comprises $N = 8{,}000$ nodes and $|E| = 56{,}400$ directed edges, resulting in an average node degree of 14.1. Edge weights encode activation probabilities, consistent with the IC diffusion mechanism, and represent the likelihood that an activated node successfully influences its neighboring nodes. The network topology and edge-weight configuration were held constant across all experimental runs to ensure controlled and comparable evaluation of influence propagation across prompting strategies and language models.

4.9.2. Activation Probability

Activation follows:

$$P(u \to v) = p_{uv}, \tag{15}$$

where p_{uv} reflects the persuasive strength of generated content, often improved under structured reasoning prompts (Hendrycks *et al.*, 2020).

4.9.3. Expected Influence Spread

The expected influence spread, denoted as $\sigma(p)$, represents the anticipated number of individuals (nodes) that become activated under a given prompting strategy p:

$$\sigma(p) = \mathbb{E}[\text{number of activated nodes}]. \tag{16}$$

This expectation is computed over multiple simulation runs of the IC model, capturing the probabilistic nature of message propagation in interconnected populations. Higher values of $\sigma(p)$ indicate that the awareness message generated under prompting strategy p is more effective at stimulating onward transmission across the network. In the context of drug prevention campaigns, an increased influence spread reflects broader educational penetration, stronger engagement across community clusters, and a greater likelihood that key health-protection messages reach individuals in diverse demographic subgroups (Bharel *et al.*, 2024).

4.9.4. Spread Gain Relative to Baseline

To quantify the improvement in dissemination attributable to each prompting strategy, the spread gain $G(p)$is defined as follows:

$$G(p) = \frac{\sigma(p) - \sigma(\text{baseline})}{\sigma(\text{baseline})} \times 100. \tag{17}$$

where $\sigma(\text{baseline})$ corresponds to the influence spread under a minimal-instruction or no-prompt condition. A positive value of $G(p)$ indicates that prompting strategy p produces a measurable enhancement in the diffusion of educational content relative to the baseline, demonstrating increased persuasive strength and communication efficacy. In drug-awareness settings, a higher spread gain signifies that the generated messages are not only more accurate but also more transmissible across social structures — facilitating accelerated reach, deeper community infiltration, and improved public health impact through more effective behavioral influence (Hendrycks *et al.*, 2020; Solaiman and Dennison, 2021).

5. Experimental Results

This section presents the empirical evaluation of the proposed framework, summarizing dataset characteristics, model performance across prompting strategies, and safety outcomes using the Edu-R, Vio-R, and MAS metrics. The results also assess how each model adapts to different user contexts and regional variations, alongside influence-spread estimates that quantify the potential real-world impact of AI-generated awareness messages.

5.1. Demographic Composition of the Social Media Corpus:

The demographic structure of the social media corpus was examined to provide an aggregate characterization of participant age and gender distributions. Age-group and gender information was derived using a privacy-preserving inference strategy based on publicly available content and platform-level indicators. When available, broad age-group cues were obtained from publicly declared date-of-birth information on platforms such as Twitter (X) and Facebook; otherwise, demographic attributes were inferred from self-disclosed linguistic cues within the content. All demographic variables were used exclusively for descriptive, aggregate-level analysis and were not employed for individual-level prediction or model training (Cesare *et al.*, 2018; Sloan *et al.*, 2015). Figure 1 summarizes the resulting age distributions. The age histogram shows that the largest proportion of participants falls within the 25–34 age group, followed by the 35–44 group, indicating a predominantly young to middle-aged population. Smaller counts appear in the 18–24 and 45–54 groups, with the lowest representation in the 55–64 category. The gender indicates that 69% of participants are male and 31% are female, reflecting an imbalanced but clearly defined gender distribution within the sample.

Figure 1. Age distributions of the study

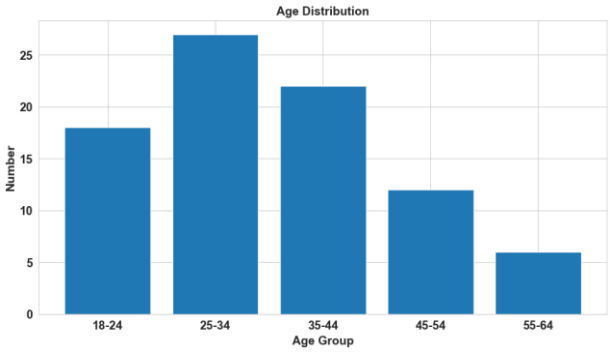
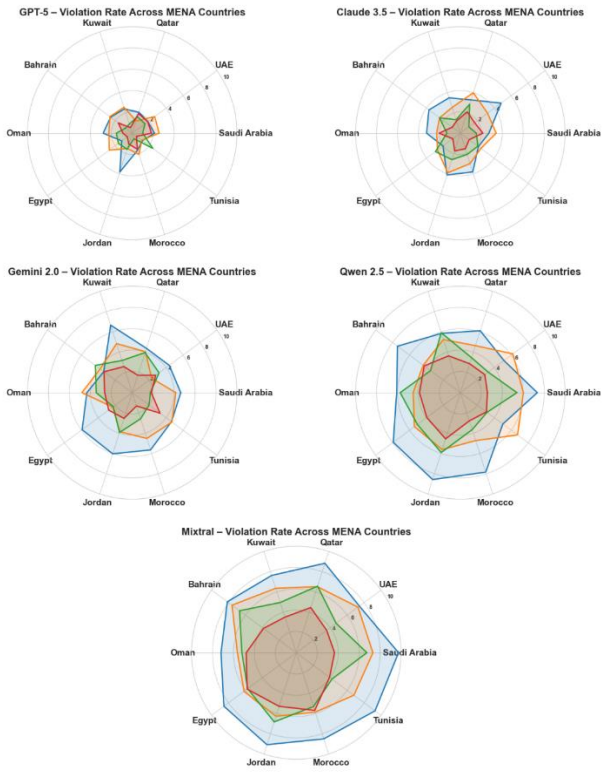


Figure 2 shows the comparative analysis of radar charts across GPT-5-mini, Claude 3.5, Gemini 2.0, Qwen 2.5, and Mixtral, demonstrating systematic differences in how LLMs respond to culturally sensitive drug-awareness prompts in the Middle East and North Africa (MENA) region. Original prompting consistently produces the highest violation rates across all models, particularly in culturally conservative Gulf countries, confirming that unstructured instructions increase the likelihood of misalignment.

Here, GPT-5-mini and Claude 3.5 have higher capacity; they exhibit the most stable and culturally grounded behavior, reflecting broader exposure to regionally relevant linguistic cues. Gemini 2.0 shows moderate robustness but greater variability across dialect-rich countries such as Egypt, Morocco, and Jordan. In contrast, Qwen 2.5 and Mixtral display markedly uneven and elevated violation profiles, with pronounced sensitivity to dialectal shifts and culturally nuanced expressions.

Enhanced prompting strategies substantially reduce violation rates across all five models. The 2NE+1SE configuration tightens distributions and mitigates country-specific spikes, particularly for Qwen and Mixtral. Further refinement with 2SE and CA+2SE produces the most uniform and minimized violation profiles, demonstrating that socio-emotional framing and campaign-adaptive guidance significantly improve alignment.

Figure 2. Vio-R distributions for GPT-5-mini, Claude 3.5, Gemini 2.0, Qwen 2.5, and Mixtral across MENA countries



5.2. Overall Classification Performance of the Evaluated LLMs:

The unified performance, as shown in Table 3, demonstrates a clear and consistent trend across all five language models: providing additional few-shot examples results in measurable and systematic improvements across Accuracy, Precision, Recall, and F1 score. Here, GPT-5-mini achieved the strongest results overall, starting from an already high baseline of 92.11% accuracy at 5-shot and reaching 95.10% at 40-shot. This upward pattern reflects GPT-5-mini's advanced ability to integrate contextual cues from minimal supervision, achieving near-optimal classification behavior with relatively few demonstrations.

Table 3. Overall performance comparison across all models

Model	Accuracy	Precision	Recall	F1 Score
GPT-5-mini	95.10%	96.22%	94.55%	95.44%
Claude	91.44%	92.33%	89.88%	91.10%
Gemini 2.0	94.20%	95.00%	93.22%	94.11%
Qwen 2.5	93.44%	94.00%	92.55%	93.22%
Mixtral	93.33%	93.77%	92.33%	93.00%

A class-level analysis of prediction behavior across the three awareness categories is presented in Figure 3. Across all evaluated models, the dominance of diagonal elements indicates stable and well-separated decision boundaries, whereas off-diagonal errors are largely confined to the High-Value and Moderately Valuable classes, reflecting their closer semantic alignment. Importantly, confusion between High-Value and Low-Value awareness remains consistently low, confirming that the observed performance gains arise from genuine discriminative capability rather than class imbalance or majority-class bias.

The ROC curves demonstrate strong discrimination performance across all evaluated models, as shown in Figure 4, with all curves positioned far above the no-skill baseline. The AUC values exceed 0.99 for every model, indicating near-perfect separation between aligned and non-aligned outputs. GPT-5-mini and Claude 3.5 achieve the highest AUCs, followed closely by Gemini 2.0, while Qwen 2.5 and Mixtral show only marginally lower values. Overall, the tightly clustered ROC curves confirm that all models exhibit robust threshold-independent discrimination, and that observed differences in safety and educational metrics arise mainly from decision-threshold effects rather than fundamental separability limitations.

Figure 3: Confusion Matrix for the Proposed Five LLM Models

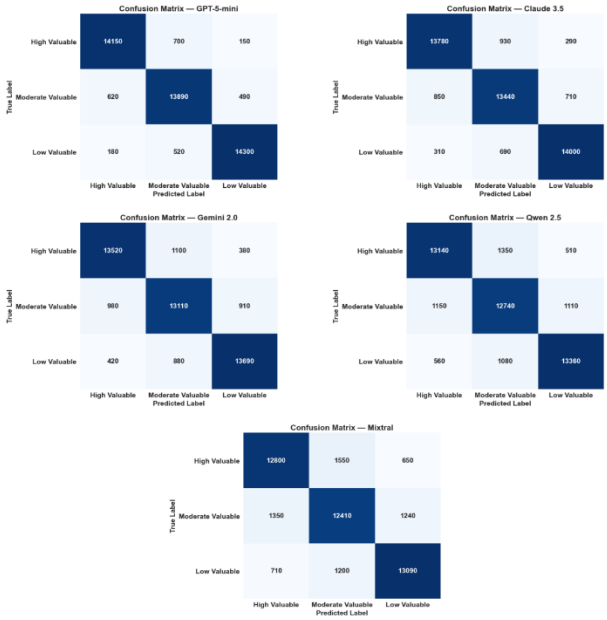
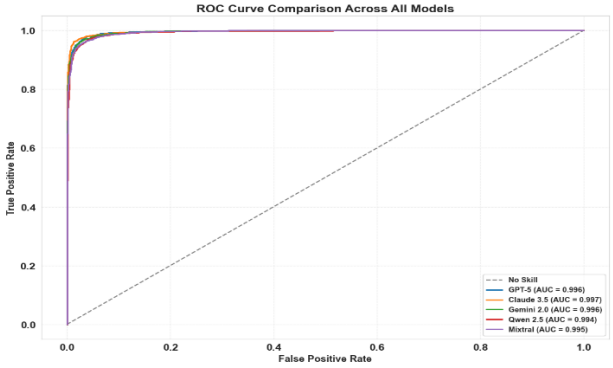


Figure 4. ROC curves for the discrimination performance of all evaluated models



5.3. Evaluating Safety and Alignment of LLMs in Drug-Awareness Messaging:

This section presents a comparative evaluation of educational quality and safety alignment across non-agent and agent-based prompting strategies using Edu-R and Vio-R. The results presented in Table 4 show that non-agent prompting achieves only moderate educational alignment, with role-based long prompts yielding Edu-R values between 80.11% and 88.72%, indicating that instruction refinement alone provides limited gains. Agent-based prompting leads to substantial improvements across all models, with 2SE and CA2NE consistently outperforming non-agent configurations. Under CA2NE_L, Edu-R increases to 95.02% for GPT-5-mini, 93.21% for Claude 3.5, and 90.55% for Gemini 2.0, with comparable upward trends for Qwen 2.5 and Mixtral. The complementary safety results presented in Table 5 demonstrate that Vio-Rs remain non-negligible under non-agent prompting (3.1%—7.7%) but decrease systematically under agent-based strategies; CA2NE_L reduces Vio-R to 1.1% for GPT-5-mini, 1.6% for Claude 3.5, and 2.1% for Gemini 2.0, confirming that structured cognitive prompting simultaneously enhances educational effectiveness and suppresses unsafe outputs in drug-awareness messaging.

Table 4. Non-agent versus agent performance across all prompting strategies for awareness Edu-R (%)

Model	Mode	Strategy	S (%)	L (%)
GPT-5-mini	Non-Agent	ORIG	78.12	83.55
		ROLE	84.40	88.72
	Agent	ORIG	81.92	86.21
		2SE	90.22	92.83
		CA2NE	93.80	95.02
Claude 3.5	Non-Agent	ORIG	75.14	81.33
		ROLE	81.60	86.92
	Agent	ORIG	79.02	84.71
		2SE	87.20	90.11
		CA2NE	91.44	93.21
Gemini 2.0	Non-Agent	ORIG	72.22	78.14
		ROLE	79.90	84.01
	Agent	ORIG	76.10	81.55
		2SE	84.44	88.01
		CA2NE	88.55	90.55
Qwen 2.5	Non-Agent	ORIG	69.18	75.40
		ROLE	76.70	81.55
	Agent	ORIG	73.44	79.20
		2SE	82.01	86.10
		CA2NE	87.05	89.44
Mixtral	Non-Agent	ORIG	67.11	73.90
		ROLE	74.44	80.11
	Agent	ORIG	71.00	77.44
		2SE	80.22	85.55
		CA2NE	85.55	88.00

Table 5. Non-agent and agent performance across all prompting strategies for Vio-R (%)

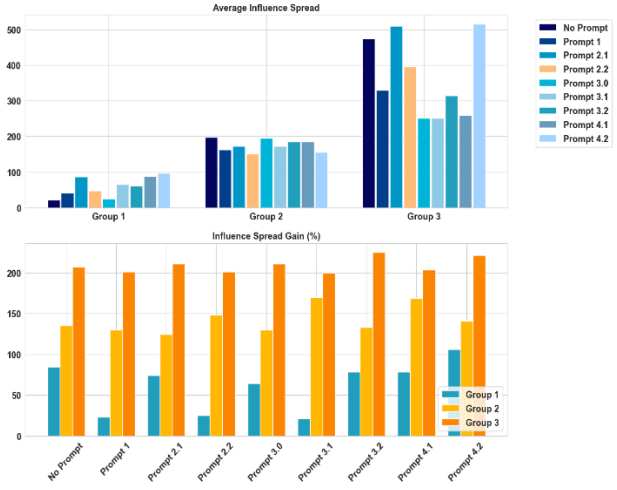
Model	Mode	Strategy	S (%)	L (%)
GPT-5-mini	Non-Agent	ORIG	4.9	4.1
		ROLE	3.55	3.1
	Agent	ORIG	3.8	3.2
		2SE	2.7	2.2
		CA2NE	1.3	1.1
Claude 3.5	Non-Agent	ORIG	6.1	5.3
		ROLE	4.6	4.1
	Agent	ORIG	4.9	4.2
		2SE	3.5	3.0
		CA2NE	1.9	1.6
Gemini 2.0	Non-Agent	ORIG	7.4	6.3
		ROLE	5.7	5.0
	Agent	ORIG	5.8	5.1
		2SE	4.2	3.7
		CA2NE	2.5	2.1
Qwen 2.5	Non-Agent		9.2	7.9
		ROLE	7.1	6.3
	Agent	ORIG	7.4	6.6
		2SE	5.6	5.0
		CA2NE	3.4	3.0
Mixtral	Non-Agent	ORIG	10.8	9.4
		ROLE	8.6	7.7
	Agent	2SE	6.8	6.1
		CA2NE	4.4	3.9

The combined evaluation of educational effectiveness, safety compliance, and annotation reliability is summarized in Table 6 using the MAS and Cohen's κ. Here, GPT-5-mini demonstrates the strongest overall performance, achieving the lowest MAS (0.0084), the highest Edu-R (95.44%), the lowest Vio-R (0.8%), and the highest agreement with human annotations (κ = 0.91), indicating highly reliable and safety-aligned awareness generation. Claude 3.5 and Gemini 2.0 maintain strong educational quality with slightly higher MAS values and moderate κ scores (0.87 and 0.84), reflecting limited residual misinformation risk. Qwen 2.5 and Mixtral exhibit higher MAS values with lower κ scores (0.80 and 0.77), indicating weaker consistency and increased exposure to misleading content. Overall, as shown in Table 6, lower MAS values consistently correspond to stronger agreement and more dependable drug-awareness communication under agent-based prompting.

Table 6. MAS and Cohen's κ for the proposed models

Model	Edu-R (%)	Vio-R (%)	MAS	Avg. κ
GPT-5-mini	95.44	0.8	0.0084	0.91
Claude 3.5	94.10	1.1	0.0117	0.87
Gemini 2.0	91.45	1.5	0.0164	0.84
Qwen 2.5	90.55	2.3	0.0254	0.80
Mixtral	89.50	3.0	0.0335	0.77

Figure 5: Influence spread of prompts and percentage-gain results



5.4. Influence-Spread Analysis Across User Groups:

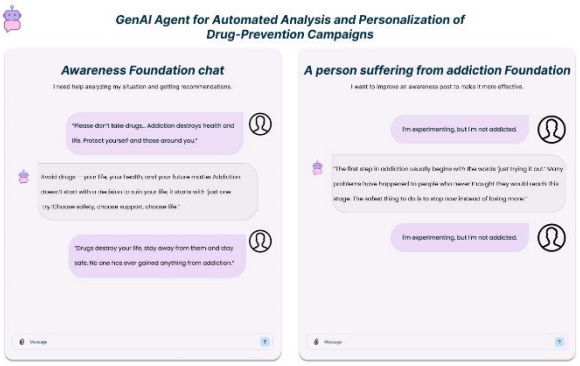
The diffusion behavior of drug-awareness messages varies systematically across user groups with different interactional and network characteristics. As shown in Figure 5, structured prompting strategies consistently increase influence spread across all groups, with the largest absolute gains observed in Group 3, reflecting its higher connectivity and stronger cascade potential. Prompting configurations that incorporate multi-stage reasoning and richer contextual framing, such as explicit sensemaking and case-based prompting, yield the strongest diffusion improvements, indicating that cognitively enriched prompts enhance node activation and accelerate message propagation relative to baseline instructions. Percentage-gain results further indicate that Groups 2 and 3 are more responsive to structured prompting than Group 1, producing disproportionately higher diffusion gains once activation probability is reinforced. This pattern aligns with influence-maximisation theory, whereby structurally cohesive and context-rich communities amplify message spread when early-stage activation is strengthened.

5.5. GenAI-Based Context-Aware Awareness Generation Across User Profiles:

The proposed GenAI agent exhibits differentiated awareness messaging behavior when interacting with users representing distinct contextual profiles. For a general awareness-oriented user, the agent produces broadly framed preventive responses that emphasize health risks, personal safety, and socially responsible decision-making. The messaging remains precautionary and non-intrusive, reinforcing awareness and risk-avoidance objectives without introducing assumptions about individual vulnerability or behavioral state.

When the interaction context reflects early-stage risk signals related to substance use, the agent adjusts both the tone and content of its responses. In this case, the generated messages move beyond generic warnings towards more targeted and supportive guidance, addressing behaviors such as experimentation and minimization while acknowledging psychological risk factors and emphasizing the importance of early interruption of harmful patterns. The contrasting interaction examples illustrated in Figure 6 demonstrate the agent's ability to modulate awareness responses according to inferred user context, enabling differentiated communication strategies for general audiences and individuals with emerging risk profiles.

Figure 6. GenAI agent interface showing adaptive responses across different user profiles



6. Conclusions

This study demonstrates the potential of next-generation GenAI models to automate, evaluate, and personalize drug prevention campaigns at scale. Leveraging a 45,000-post multi-platform corpus and a structured prompting framework spanning Original, Role-Based, 2SE, and CA2NE variants, the proposed GenAI agent generates reliable and context-aware outputs for drug-awareness messaging. Among the five LLMs, GPT-5-mini achieves the strongest overall performance, attaining 95.10% accuracy, 96.22% precision, 94.55% recall, and a 95.44% F1 score, consistently outperforming Claude 3.5, Gemini 2.0, Qwen 2.5, and Mixtral.

Educational alignment improves consistently with agent-based prompting; GPT-5-mini's Edu-R increases from 78.12% under minimal instructions to over 95% under long CA2NE prompting, with Claude 3.5 and Gemini 2.0 exhibiting comparable directional gains. Safety analysis further confirms the effectiveness of structured cognitive prompting: across all models, Vio-Rs are reduced to low-single-digit values, corresponding to approximately 96%–99% safety-aligned outputs, compared with substantially higher violation exposure under non-agent prompting. Discrimination analysis based on ROC curves yields AUC values ranging from 99.4% to 99.7%, indicating strong threshold-independent separability between aligned and non-aligned awareness content.

Agreement analysis using Cohen's κ further supports the robustness of the proposed framework, with high κ scores indicating strong consistency between model predictions and ground-truth awareness labels across agent-based configurations. These agreement results complement accuracy-based metrics by confirming that performance gains reflect genuine alignment rather than chance agreement. Influence-spread simulations additionally show that cognitively rich prompting strategies amplify diffusion effectiveness, with the largest percentage gains observed in highly connected demographic clusters and youth-driven networks.

Collectively, these findings establish that a GenAI agent powered by structured prompting provides a robust, transparent, and scalable mechanism for real-time evaluation and optimization of digital drug prevention campaigns. The unified evaluation framework integrating Edu-R, Vio-R, MAS, ROC–AUC, Cohen's κ, and diffusion modeling offers policymakers and public health organizations an evidence-driven foundation for designing safer, more targeted, and culturally responsive awareness strategies.

Future work should extend this framework to multimodal awareness content (e.g. images and short-form video) and evaluate real-world deployment to assess operational robustness, responsiveness to emerging trends, and longer-term behavioral impact.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author.

Acknowledgement

This work was supported by the Bashaer Society.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU253464].

Conflict of Interest

The authors declare no conflict of interest.

Biographies

Mohammed Al Jaafari

Department of Management Information Systems, School of Business, King Faisal University, Al-Ahsa, 31982, Saudi Arabia, 009660557122210, mahaljaafari@kfu.edu.sa

Dr. Aljaafari is a Saudi Assistant Professor of Information Systems, specializing in Computer Information Systems. He received the master's degree from Nova Southeastern University, USA, and the Ph.D. degree from Curtin University, Australia. His research interests include social commerce, artificial intelligence, data analytics, and emerging digital technologies. He has published scholarly research in high-impact international journals. He serves as Vice Dean for Academic Affairs and has received the Chancellor's Letters of Commendation from Curtin University, Australia, in recognition of his academic excellence and contributions to higher education.

ORCID: 0000-0002-1151-9217

Shaymaa E. Sorour

Department of Management Information Systems, School of Business, King Faisal University, Al-Ahsa, 31982, Saudi Arabia, 009660544315220, ssorour@kfu.edu.sa

Sorour is an Egyptian Associate Professor of Computer Science, specializing in artificial intelligence. She received the Ph.D. degree in Computer Science from Kyushu University, Japan. Her research focuses on the development of data-driven intelligent decision-support systems, with application areas including smart irrigation management, intelligent sensing, and AI-enabled sustainability solutions. She has published extensively in high-quality, peer-reviewed international journals and actively contributes to the scientific community through professional service as a reviewer and editorial board member for leading journals in artificial intelligence and data science.

ORCID: 0000-0003-0805-2705

References

Ahmad, M., Batyrshin, I. and Sidorov, G. (2025). Sentiment analysis using a large language model–based approach to detect opioids mixed with other substances via social media: Method development and validation. *JMIR Infodemiology*, **5**(n/a), e70525. DOI:10.2196/70525

Alamoodi, A.H., Zaidan, B.B., Zaidan, A.A., Zaidan, A.A., Albahri, O.S., Mohammed, K.I., Malik, R.Q., Almahdi, E.M., Chyad, M.A., Tareq, Z.,Albahri, A.S., Hameed, H. and Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, **167**(114155). DOI:10.1016/j.eswa.2020.114155

Albarrak, K.M. and Sorour, S.E. (2024). Boosting institutional identity on X using NLP and sentiment analysis: King Faisal University as a case study. *Mathematics*, **12**(12), 1806. DOI:10.3390/math12121806

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I. and Precise4Q Consortium (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, **20**(1), 310. DOI:10.1186/s12911-020-01332-6

Bandeira, A., Goncalves, L.H., Holl, F., Shaibu, I.U., Goncalves, M.L., Pavinda, R., Paudel, S., Berionni, A., WFPHA, Y., Purnat, T.D. and Mackey, T. (2025). Viewpoint on the intersection among health information, misinformation, and generative AI technologies. *JMIR Infodemiology*, **5**(1), e69474. DOI:10.2196/69474

Brandao, B.M. and Denny, B.T. (2024). What instagram means to me: Links between social anxiety, instagram contingent self-worth, and automated textual analysis of linguistic authenticity. *Affective Science*, **5**(4), 449–57. DOI:10.1007/s42761-024-00267-9

Bharel, M., Auerbach, I., Nguyen, V. and DeSalvo, K.B. (2024). Transforming public health practice with generative artificial intelligence: Article examines how generative artificial intelligence could be used to transform public health practice in the US. *Health Affairs*, **43**(6), 776–82. DOI:10.1377/hlthaff.2024.00050

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, **33**(n/a), 1877–901.

Cesare, N., Lee, H., McCormick, T., Spiro, E. and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography*, **55**(5), 1979–99. DOI:10.1007/s13524-018-0715-2

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46. DOI:10.1177/001316446002000104

Deng, T., Urbaczewski, A., Lee, Y.I., Barman-Adhikari, A. and Dewri, R. (2024). Identifying Marijuana Use Behaviors Among Youth Experiencing Homelessness Using a Machine Learning–Based Framework: Development and Evaluation Study. *JMIR AI*, **3**(1), e53488. DOI:10.2196/53488

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. Cambridge: MIT Press.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D. and Steinhardt, J. (2020). *Aligning AI With Shared Human Values*. Available at: https://arxiv.org/abs/2008.02275 (accessed on 10/11/2025).

Khakpaki, A. and Sepehri, H. (2025). AI in addiction: Harnessing technology for diagnosis, prevention, and recovery: A narrative review. *Addiction and Substance Abuse*, **3**(1), 1–7. DOI:10.46439/addiction.3.008

Khosravi, M., Zare, Z., Mojtabaeian, S.M. and Izadi, R. (2024). Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews. *Health Services Research and Managerial Epidemiology*, **11**(n/a), 1–13. DOI:10.1177/23333928241234863

Kojima, T., Gu, S., Reid, M., Matsuo, Y. and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, **35**(n/a), 22199–213.

Krippendorff, K. (2018). *Content analysis: An Introduction to its Methodology*. California: Sage Publications.

Lendvai, G.F. (2025). Reddit in scholarly reception: A bibliometric assessment of the front page of the internet. *Quality and Quantity*, **n/a**(n/a), 1–27. DOI:10.1007/s11135-025-02416-z

Li, W., Hua, Y., Zhou, P., Zhou, L., Xu, X. and Yang, J. (2025). Characterizing public sentiments and drug interactions in the COVID-19 pandemic using social media: Natural language processing and network analysis. *Journal of Medical Internet Research*, **27**(n/a), e63755. DOI:10.2196/63755

Maharjan, J., Zhu, J., King, J., Phan, N., Kenne, D. and Jin, R. (2025). Large-scale deep learning–enabled infodemiological analysis of substance use patterns on social media: Insights from the COVID-19 pandemic. *JMIR Infodemiology*, **5**(n/a), e59076.

Manning, C., and Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT press.

Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y. and Cheungpasitporn, W. (2024). Chain of thought utilization in large language models and application in nephrology. *Medicina*, **60**(148), 1–19.

Nasser, B.S.A. and Abu-Naser, S.S. (2024). Artificial intelligence in digital media: Opportunities, challenges, and future directions. *International Journal of Academic and Applied Research (IJAAR)*, **8**(6), 1–10.

Nishan, M.N.H. (2025). AI-powered drug discovery for neglected diseases: Accelerating public health solutions in the developing world. *Journal of Global Health*, **15**(n/a), 03002. DOI:10.7189/jogh.15.03002

Nwanakwaugwu, A.C., Andrew-Vitalis, N., Kwakpovwe, P., Emakporuena, D. and Eboesomi, E. (2025). Personalizing medicine for fake drug prevention with Ai-driven digital twins. *AI-Powered Digital Twins for Predictive Healthcare: Creating Virtual Replicas of Humans*, **n/a**(n/a), 325–58 DOI:10.4018/979-8-3373-0538-7.ch010

Olawade, D.B., Wada, O.J., David-Olawade, A.C., Kunonga, E., Abaire, O. and Ling, J. (2023). Using artificial intelligence to improve public health: A narrative review. *Frontiers in Public Health*, **11**(n/a), 1-9. DOI:10.3389/fpubh.2023.1196397

Olivares-De la Fuente, P., Jiménez-García, E., Y. and García-López, Ó. (2025). Twitter and YouTube as digital tools in higher education: A systematic review. *Frontiers in Education*, **10**(n/a), 1–9. DOI:10.3389/feduc.2025.1625803

Panteli, D., Adib, K., Buttigieg, S., Goiana-da-Silva, F., Ladewig, K., Azzopardi-Muscat, N., Figueras, J., Novillo-Ortiz, D. and McKee, M. (2025). Artificial intelligence in public health: Promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet Public Health*, **10**(5), e428–32. DOI:10.1016/S2468-2667(25)00036-2

Plackett, R., Steward, J.M., Kassianos, A.P., Duenger, M., Schartau, P., Sheringham, J., Cooper, S., Biddle, L., Kidger, J. and Walters, K.(2025). The effectiveness of social media campaigns in improving knowledge and attitudes toward mental health and help-seeking in high-income countries: Scoping review. *Journal of Medical Internet Research*, **27**(n/a),1–17. DOI:10.2196/68124

Sloan, L., Morgan, J., Burnap, P. and Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one*, **10**(3), e0115545. DOI:10.1371/journal.pone.0115545

Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, **34**(n/a), 5861–73.

Sorour, S.E. and Almusallam, N. (2025). L3D-RAG: Leveraging LLaMA 3.1 and DeepSeek for Reddit analysis. *Alexandria Engineering Journal*, **131**(n/a), 125–52. DOI:10.1016/j.aej.2025.09.070

Uddin, J., Feng, C. and Xu, J. (2025). Health communication on the internet: Promoting public health and exploring disparities in the generative AI era. *Journal of Medical Internet Research*, **27**(n/a), e66032. DOI:10.2196/66032

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, **30**(n/a),1–11.

Villanueva-Miranda, I., Xie, Y. and Xiao, G. (2025). Sentiment analysis in public health: A systematic review of the current state, challenges, and future directions. *Frontiers in Public Health*, **13** (n/a),1–23. DOI:10.3389/fpubh.2025.1609749

Ye, Y., Pandey, A., Bawden, C., Sumsuzzman, D. M., Rajput, R., Shoukat, A., Singer, B.H., Moghadas, S.M. and Galvani, A.P. (2025). Integrating artificial intelligence with mechanistic epidemiological modeling: A scoping review of opportunities and challenges. *Nature Communications*, **16**(581), 1–18. DOI:10.1038/s41467-024-55461-x

Zhao, Z., Wu, J., Li, T., Sun, C., Yan, R. and Chen, X. (2021). Challenges and opportunities of AI-enabled monitoring, diagnosis and prognosis: A review. *Chinese Journal of Mechanical Engineering*, **34**(56), 1–29. DOI:10.1186/s10033-021-00570-7

Zhu, N., Zhao, F., Wang, L., Ding, R. and Xu, T. (2022). A discrete learning fruit fly algorithm based on knowledge for the distributed no-wait flow shop scheduling with due windows. *Expert Systems with Applications*, **198**(116921),1–18. DOI:10.1016/j.eswa.2022.116921

Copyright