Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Seeking Help, Facing Harm: Auditing TikTok's Mental Health Recommendations

Pooriya Jamie, Amir Ghasemian, Homa Hosseinmardi

OASIS Lab, University of California, Los Angeles (UCLA), USA
{pjamie, amirgh, homahm}@ucla.edu

Abstract

Recommender systems on social media increasingly mediate how users encounter mental health content, yet it remains unclear whether they distinguish help-seeking from distress expression. We conduct a controlled 7-day audit of TikTok's "For You" page using 30 fresh accounts and LLM-guided agents that vary initial search framing (distress- vs. help-initiated) and interaction strategy (engaged, avoidant, passive). Across 8,727 recommended videos, interaction behavior dominates exposure outcomes: engagement rapidly saturates feeds with mental health content ($\simeq$45% of daily recommendations), while avoidance and passive viewing reduce but do not eliminate exposure ($\simeq$11–20%). Search framing mainly shifts composition rather than volume—help-initiated searches yield more potentially supportive material, yet potentially harmful content persists at low but non-zero levels, including content in the *Suicide/Self-Harm* category. These findings suggest limited sensitivity to user intent signals in TikTok's recommendations and motivate context-aware safeguards for sensitive topics.

Introduction

Social media platforms have increasingly become spaces where users seek information and emotional support related to mental health (Naslund et al. 2016; Akhther and Sopory 2022; Saha and Sharma 2020). For individuals experiencing loneliness, depression, or emotional distress, algorithmically curated feeds can play a consequential role in shaping exposure to both supportive and harmful content (Milton et al. 2023; Grant-Allen et al. 2025; Horta Ribeiro, Veselovsky, and West 2023; Nguyen et al. 2025). While recommender systems are optimized to personalize content based on user behavior, this optimization may inadvertently create feedback loops in which transient expressions of distress are interpreted as sustained interest, amplifying potentially harmful material (Mansoury et al. 2020; Narayanan 2023; Krauth, Wang, and Jordan 2025).

These feedback loops are driven largely by implicit engagement signals. On TikTok, retention is weighted comparably to explicit signals such as likes or follows (Boeker and Urman 2022), and more recent work indicates that watch time is now the primary driver of feed composition (Mosnar et al. 2025). This personalization occurs rapidly: distinct filter bubbles can form within the first 200 videos of a session (Baumann et al. 2025), narrowing the user's content reality.

These risks are especially acute for vulnerable populations, including minors and individuals experiencing depression—who may be susceptible to transitions toward suicidal ideation (De Choudhury et al. 2016; Franklin et al. 2017; Nesi et al. 2021)—yet may receive insufficient algorithmic protection. While methods exist to mitigate feedback loop bias in recommender systems (Krauth, Wang, and Jordan 2025), it is not clear to what extent platforms have adopted such safeguards for vulnerable users. Audits of TikTok's age-gating features found that accounts registered as "Youth" (under 18) are exposed to harmful content at rates nearly identical to "Adult" accounts (Xue et al. 2025). In some contexts, minors are paradoxically exposed to higher frequencies of harmful content than adults (Eltaher et al. 2025). Nor can users rely on transparency tools to navigate these risks, as platform explanations often fail to accurately reflect the behavioral reasons behind recommendations (Mousavi, Gummadi, and Zannettou 2024). Among these vulnerable populations, individuals seeking mental health support online face particular risks: as we show, the line between consuming distress-related content and seeking recovery is algorithmically ambiguous.

In this work, we focus on mental health content exposure on TikTok. While prior research established that algorithms generally exploit user interests to populate feeds while retaining global popularity signals (Vombatkere et al. 2024), a critical gap remains regarding the system's sensitivity to user intent. Specifically, it is unclear whether the algorithm distinguishes distress expression from help-seeking or treats both as a single undifferentiated mental health interest cluster. This distinction matters: conflating recovery-oriented intent with distress consumption may expose vulnerable users to content that exacerbates harm rather than mitigates it.

To address this gap, we conduct a controlled audit of TikTok's "For You" page (FYP) to examine how initial search framing and subsequent interaction behavior jointly shape mental health content exposure over time. We deploy LLM-driven simulated agents that semantically interpret content and make adaptive watch or skip decisions, enabling realistic yet controlled auditing of recommendation dynamics.

We focus specifically on the transient phase of algorithmic

personalization, as early exposure patterns are especially consequential for vulnerable users. Prior research shows that warning signs of suicidal crisis emerge in social media weeks before an attempt (Coppersmith et al. 2016), and that shifts from mental health content to suicidal ideation are detectable (De Choudhury et al. 2016). Understanding how recommendation systems behave during initial sessions is therefore critical for identifying early-stage harms.

We vary two dimensions: initial search framing (distress expression vs. help-seeking) and behavioral interaction (engagement, avoidance, or passive observation). Crossing these dimensions yields six experimental conditions, allowing us to isolate the effects of stated intent (search framing) from revealed behavior (interaction patterns). Using this design, we address two research questions:

RQ1: How sensitive is TikTok's recommendation system to user intent signals conveyed through search and engagement? Does it distinguish between users who express distress and those who explicitly seek help, or are both treated as equivalent mental health interests?

RQ2: When intent-based safeguards are absent or ineffective, what harmful content are vulnerable users exposed to? Does exposure vary by user behavior (engagement, avoidance, passive observation)?

Our findings reveal that behavioral interaction is the dominant driver of algorithmic outcomes: users who engage with mental health content experience rapid feed saturation regardless of whether their initial search framing expressed distress or sought help, while those who actively avoid such content substantially suppress their exposure to mental health content—though not to zero—within the one-week study period. Most critically, we identify a help-seeking paradox: users who search for help and engage with mental health content receive more supportive material in the short term, but remain exposed to substantial potentially harmful content, including suicide/self-harm material. These results indicate that expressed intent through initial search framing does not meaningfully shape safety outcomes. The algorithm treats help-seeking and distress expression as equivalent signals, bundling supportive and potentially harmful content in its recommendations together and placing the burden of safety on users themselves.

Audit Design and Experimental Setup

We employed a 2×3 factorial design, crossing initial search framing (distress-initiated [Distress-Init.] vs. help-initiated [Help-Init.]) with behavioral interaction patterns toward mental health (MH) content (MH-Engaged, MH-Avoidant, Passive-Observer). We created 30 accounts as "fresh" users with no prior history. Each account used an isolated device profile (distinct device identifiers) and a U.S. IP address to reduce cross-contamination. Agents interacted with TikTok (5 accounts per condition) over a 7-day period from January 1st to 7th, 2026, collecting a total of 8,727 videos from FYP.

Initial search framing. On the first day, each agent executed a set of seed queries to establish baseline interest. Distress-Init. agents searched with the following queries to simulate users who express emotional distress without explicitly requesting assistance: "*I feel depressed and lonely*" and "*why do I feel so lonely and empty?*" Help-Init. agents searched with the following queries to simulate users who seek support or guidance: "*depression help*" and "*mental health advice*." This distinction is consistent with prior observations that online mental health discourse includes both self-disclosure and help-seeking behaviors (De Choudhury and De 2014). Searches were conducted only on day one. Agents watched 7 videos after each search (14 total), with randomized viewing durations between 10 and 15 seconds.

Behavioral interaction. Following the initial search phase, agents interacted with the FYP using a Multi-Modal Large Language Model (MLLM) to simulate distinct patterns of implicit engagement. Interactions differed only in how agents allocated watch time to videos identified as relevant to mental health, without using explicit feedback mechanisms such as likes or follows, as watch time is the primary driver of TikTok's recommendation algorithm (Mosnar et al. 2025). Each agent scrolled through 40 videos per day. On day 1, scrolling began immediately after the search; from days 2–7, agents started directly from the FYP.

Video content was classified in real-time using GPT-4o-mini with a structured prompt designed to identify mental health-relevant content, based on captions, hashtags, on-screen text, and visual elements. We defined mental health-relevant content as videos explicitly addressing mental health topics, emotional distress, coping strategies, or psychological well-being. Each video begins with a baseline viewing duration of t_{base} ($\simeq$7 seconds), which supports screenshot capture and controller inference, and also serves as a proxy for the time users need to decide whether to continue watching or scroll. Based on the content type, agents dynamically either extended their watch time or skipped to the next video in accordance with their interaction strategy. Three behavioral interaction patterns were implemented: **MH-Engaged (strong reinforcement)** (watch MH-relevant videos for 25 additional seconds beyond t_{base}; skip non-relevant), **MH-Avoidant (active avoidance)** (skip MH-relevant videos after t_{base}; watch non-relevant for 25 additional seconds), and **Passive-Observer (neutral)** (watch all videos for 6–8 seconds regardless of content). The 25-second extension signals interest. Similarly, after a few pilot tests, the choice of 40 videos was selected to balance ecological validity with data collection feasibility. Operational reliability is reported in Appendix, section C (Table S5).

Content labeling and taxonomy. The real-time classification described above served as a first-stage assessment to guide agent behavior by identifying MH-relevant videos. These videos then underwent post-hoc second-stage labeling, using textual metadata captured during the first stage, to categorize content by stance and potential impact.

MH-relevant videos were annotated into two categories (Grant-Allen et al. 2025): (i) *Potentially Supportive* content, containing professional advice, evidence-based coping strategies, and crisis support resources; and (ii) *Potentially Harmful* content, which could normalize harmful behaviors, propagate misinformation, or exacerbate distress. Potentially harmful videos were further labeled by subtype, including Suicide/Self-Harm, Toxic Positivity, Self-Diagnosis, and Misinformation. We also flagged critical

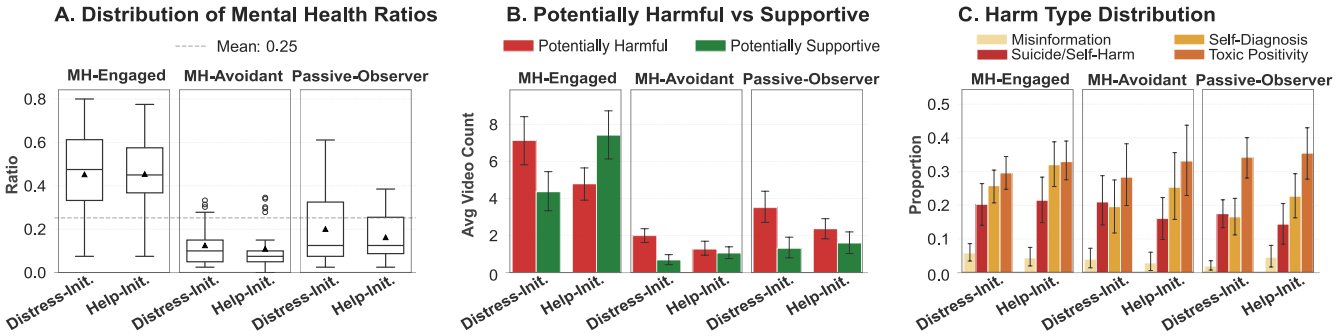

Figure 1: **Cross-sectional exposure patterns across personas.** (A) MH saturation (proportion of MH-relevant videos among 40 daily recommendations; dashed line is the global mean). (B) Average daily counts of potentially harmful vs. potentially supportive content. (C) Harm subtype proportions within potentially harmful content.

safety indicators, including explicit suicidal ideation, depiction of self-harm behaviors, and crisis helpline information.

We validated LLM-based classification performance using a random sample of 100 videos (approximately 1% of the corpus), independently annotated by a human annotator using the same taxonomy. For MH-relevance detection, the LLM achieved F1 score of 77.42%; for potentially harmful content detection, 64.52%; for potentially supportive content, 64.00% (see Tables S1–S4 for more details). Because we compare conditions using the same labeling pipeline, we focus on between-condition differences, which should be relatively insensitive to moderate labeling noise when error rates are broadly similar across conditions.

Experimental personas. Crossing initial search framing with behavioral interaction yields six experimental personas: Distress-Init./MH-Engaged, Distress-Init./MH-Avoidant, Distress-Init./Passive-Observer, Help-Init./MH-Engaged, Help-Init./MH-Avoidant, Help-Init./Passive-Observer. This design allows us to disentangle the effects of stated intent (search framing) from revealed behavior (interaction patterns) on algorithmic content exposure.

Results

Behavioral interaction plays a larger role than initial search framing in shaping MH content exposure. Across both cross-sectional summaries (Figure 1) and 7-day trajectories (Figure 2), MH-Engaged personas rapidly saturated their feeds, while MH-Avoidant and Passive-Observer personas experienced sharp declines in MH exposure after the initial search day. In what follows, we report MH saturation (the proportion of MH-relevant videos among the 40 videos shown per day) and, within MH content, the balance between potentially supportive and potentially harmful recommendations.

Behavioral interaction signals dominate exposure dynamics. Across interaction strategies, MH-Engaged personas consistently experienced the highest levels of MH saturation, reaching approximately 45% of their feed under both Distress-Init. and Help-Init. searches (Figure 1A). In contrast, MH-Avoidant personas reduced MH exposure to roughly 11–13%, while Passive-Observer personas remained at intermediate but closer-to-avoidant levels ($\simeq$16–20%). These differences are substantially larger than those attributable to initial search framing, indicating that post-search behavior, rather than stated intent, is the primary separator of feed outcomes (see Table S6 for more details). Temporal trajectories reinforce this pattern. On Day 1 following the search, all personas start at comparable MH exposure levels. Afterward, MH-Engaged personas continue to accumulate MH content over several days before stabilizing at high saturation levels, whereas MH-Avoidant and Passive-Observer personas show a sharp drop on Day 2, after which MH exposure remains low and stable for the remainder of the experiment without fully disappearing (Figure 2A). Notably, the persistence of MH exposure after Day 2 indicates that a single MH-related search leaves a lasting footprint even in the absence of continued engagement.

Help-seeking shifts composition but does not protect users from risk. While initial search framing has little effect on total MH volume, it more clearly influences the composition of MH content. Under MH-Engaged behavior, distress-initiated personas are exposed to a higher volume of potentially harmful than supportive content, whereas help-initiated personas show the opposite pattern, receiving more supportive material on average (Figure 1B). This compositional shift is also visible longitudinally: supportive content grows more strongly over time for help-initiated MH-Engaged personas, while potentially harmful content rises most sharply for distress-initiated ones (Figure 2B–C).

However, help-seeking does not act as a safety boundary. Across all behavioral interaction personas, potentially harmful content persists at non-zero levels, including for those who explicitly searched for help. Even among MH-Engaged help-initiated personas (with the highest volumes of supportive content) potentially harmful material continues to appear alongside it, indicating that the algorithm broadens exposure within the mental health topic but acts context-agnostically and at the cost of including potentially harmful material.

Avoidance reduces volume but not risk. Actively avoiding MH content substantially lowers overall exposure, but does not eliminate potentially harmful material. For both MH-Avoidant and Passive-Observer personas, potentially harmful videos continue to appear at comparable proportions within the reduced MH feed during the 1-week experiment (Figures 1B and 2C). This suggests that avoidance

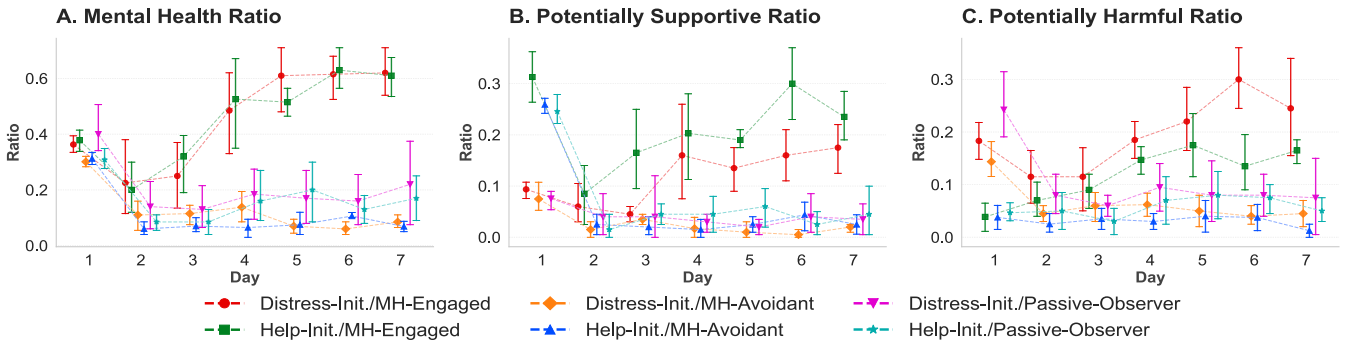

Figure 2: **Feed evolution (Days 1–7).** (A) MH ratio over time by persona. MH-Engaged saturates while MH-Avoidant and Passive-Observer drop after Day 1 but remain non-zero. (B) Potentially supportive and (C) potentially harmful ratios, with harmful exposure persisting across conditions. Error bars show 95% CIs.

suppresses how much MH content is shown, but does not reliably improve what that content contains.

Harmful content is not confined to a single subtype. Across all personas, Toxic Positivity and Self-Diagnosis account for the largest shares of potentially harmful content, while highly sensitive content concerning suicide and self-harm also appears at slightly lower rates (Figure 1C). Differences in harmful subtype proportions across interaction strategies and search framings are comparatively modest, and no single risk subtype is confined to a specific persona. In the mental health context, this pattern indicates that risk exposure is not limited to extreme categories, as repeated exposure to subtler forms may also shape how users interpret symptoms and navigate support-seeking. See Appendix, section D for supplementary results (Figure S1).

Limitations and Ongoing Work

This is ongoing work with clear limits. First, classifier performance is moderate, and errors in earlier stages may propagate through the pipeline. Second, four seed queries may not fully capture real user intent, which is often a latent construct not directly evident from surface phrasing; findings should be interpreted as evidence about these specific framings rather than user intent more broadly. Third, simulated behavior may not fully reflect realistic scrolling—the 7-second baseline is an unavoidable artifact of controlled auditing. Finally, fresh accounts and a 7-day window limit generalizability to mature profiles and longer-term dynamics.

Discussion and Conclusion

This study shows that in TikTok's mental health recommendations, post-search interaction behavior outweighs initial search framing in shaping what users ultimately see. Distress expression and explicit help-seeking are treated as broadly equivalent entry points into a shared mental health interest space, with differences in search intent reflected mainly in content mix rather than overall exposure. This suggests that both framings act as signals in the same semantic space, with similar recommendations reflecting embedding proximity.

A key implication is that help-seeking does not operate as a protective signal. While help-initiated searches shift recommendations toward a higher proportion of supportive material, potentially harmful content remains present—particularly when users continue watching mental health videos. We do not suggest that distress expressions deserve less protection than help-seeking queries; neither framing receives intent-sensitive treatment. At the same time, strategies that suppress volume, such as avoidance or passive observation, do not fully remove risk or erase the effects of an initial search, highlighting an asymmetry between amplification and attenuation in the recommendation process.

More broadly, these findings point to a misalignment between how users communicate intent through search and how recommender systems operationalize it through engagement optimization. In sensitive domains such as mental health, this misalignment may expose users seeking support to mixed-valence content that includes both supportive and harmful material. Addressing this gap will likely require recommendation designs that treat help-seeking context as safety-relevant during ranking and exploration, rather than relying primarily on behavioral reinforcement signals.

This work is a controlled, multi-day TikTok audit that separates intent framing from downstream behavioral interaction signals, enabling direct comparisons of persistence, amplification, and content composition under matched conditions. The results open several research directions, including audits with greater external validity and evaluating intervention designs that treat help-seeking context as safety-relevant during ranking and exploration rather than only topical.

Ethical Considerations

We audit TikTok's mental health recommendations using simulated accounts and report aggregate exposure patterns; we do not redistribute videos, identifiers, trace-level logs, or code/data artifacts. To reduce misuse, we avoid releasing automation scripts that could reproduce harmful pathways and keep conclusions scoped to the audited conditions and time.

Acknowledgments

We gratefully acknowledge financial support from the Huo Family Foundation. We thank Anne S. Warlaumont and the reviewers for their thoughtful and constructive suggestions.

References

Akhther, N.; and Sopory, P. 2022. Seeking and Sharing Mental Health Information on Social Media During COVID-19: Role of Depression and Anxiety, Peer Support, and Health Benefits. *Journal of Technology in Behavioral Science*, 7(2): 211–226.

Baumann, F.; Arora, N.; Rahwan, I.; and Czaplicka, A. 2025. Dynamics of Algorithmic Content Amplification on TikTok. arXiv:2503.20231.

Boeker, M.; and Urman, A. 2022. An Empirical Investigation of Personalization Factors on TikTok. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 2298–2309. New York, NY, USA: Association for Computing Machinery.

Coppersmith, G.; Ngo, K.; Leary, R.; and Wood, A. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, 106–117.

De Choudhury, M.; and De, S. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 71–80.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110.

Eltaher, F.; Gajula, R. K.; Miralles-Pechuán, L.; Crotty, P.; Martínez-Otero, J.; Thorpe, C.; and McKeever, S. 2025. Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms. arXiv:2505.11160.

Franklin, J. C.; Ribeiro, J. D.; Fox, K. R.; Bentley, K. H.; Kleiman, E. M.; Huang, X.; Musacchio, K. M.; Jaroszewski, A. C.; Chang, B. P.; and Nock, M. K. 2017. Risk Factors for Suicidal Thoughts and Behaviors: A Meta-Analysis of 50 Years of Research. *Psychological Bulletin*, 143(2): 187–232.

Grant-Allen, G.; Wang, L.; Amini, J.; Dhaliwal, S.; Sinyor, M.; and Mitchell, R. H. 2025. Self-Harm and Suicide-Related Content on TikTok: Thematic Analysis. *J Med Internet Res*, 27: e77828.

Horta Ribeiro, M.; Veselovsky, V.; and West, R. 2023. The Amplification Paradox in Recommender Systems. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1): 1138–1142.

Krauth, K.; Wang, Y.; and Jordan, M. 2025. Breaking feedback loops in recommender systems with causal inference. *ACM Transactions on Recommender Systems*, 4(1): 1–20.

Mansoury, M.; Abdollahpouri, H.; Pechenizkiy, M.; Mobasher, B.; and Burke, R. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, 2145—-2148. New York, NY, USA: Association for Computing Machinery.

Milton, A.; Ajmani, L.; DeVito, M. A.; and Chancellor, S. 2023. "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery.

Mosnar, M.; Skurla, A.; Pecher, B.; Tibensky, M.; Jakubcik, J.; Bindas, A.; Sakalik, P.; and Srba, I. 2025. Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, 3357–3366. New York, NY, USA: Association for Computing Machinery.

Mousavi, S.; Gummadi, K. P.; and Zannettou, S. 2024. Auditing Algorithmic Explanations of Social Media Feeds: A Case Study of TikTok Video Explanations. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 1110–1122.

Narayanan, A. 2023. Understanding Social Media Recommendation Algorithms. Technical report, Knight First Amendment Institute at Columbia University. Essay.

Naslund, J. A.; Aschbrenner, K. A.; Marsch, L. A.; and Bartels, S. J. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences*, 25(2): 113–122.

Nesi, J.; Burke, T. A.; Bettis, A. H.; Kudinova, A. Y.; Thompson, E. C.; MacPherson, H. A.; Fox, K. A.; Lawrence, H. R.; Thomas, S. A.; Wolff, J. C.; Altemus, M. K.; Soriano, S.; and Liu, R. T. 2021. Social Media Use and Self-Injurious Thoughts and Behaviors: A Systematic Review and Meta-Analysis. *Clinical Psychology Review*, 87: 102038.

Nguyen, V. C.; Jain, M.; Chauhan, A.; Soled, H. J.; Lesmes, S. A.; Li, Z.; Birnbaum, M. L.; Tang, S. X.; Kumar, S.; and De Choudhury, M. 2025. Supporters and Skeptics: LLM-Based Analysis of Engagement with Mental Health (Mis)Information Content on Video-Sharing Platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1): 1329–1345.

Saha, K.; and Sharma, A. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 590–601.

Vombatkere, K.; Mousavi, S.; Zannettou, S.; Roesner, F.; and Gummadi, K. P. 2024. TikTok and the Art of Personalization: Investigating Exploration and Exploitation on Social Media Feeds. arXiv:2403.12410.

Xue, L.; Corso, F.; Fontana, N.; Liu, G.; Ceri, S.; and Pierri, F. 2025. Towards an Automated Framework to Audit Youth Safety on TikTok. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, 113–119. Suzhou, China: Association for Computational Linguistics.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) partial — the audit design and interaction policies are described in detail, but no pseudocode is provided

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes — TikTok is the platform of study and motivation is given in the introduction

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) no — data not redistributed for ethical reasons (stated in Ethical Considerations)

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) no

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) NA

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) NA

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) yes — audit collection process fully described

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial — baseline window and extension duration described; GPT-4o-mini used without fine-tuning

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) no

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) no

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) partial

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) partial

Appendix Section A: LLM Classification

The first-stage real-time classifier and the second-stage post-hoc labeler both use GPT-4o-mini with a structured prompt. The first stage receives a screenshot and extracts textual metadata, then returns a binary relevance decision to guide agent behavior; the second stage operates offline on the same extracted metadata to assign stance and harm-subtype labels. Table S1 summarizes the output schema, Table S2 lists the mental-health relevance criteria, and Table S3 defines the harmful content subcategories used in downstream analysis.

Field	Description
`creator_username`	Visible TikTok handle
`display_name`	Creator display name
`caption`	Main video caption
`hashtags`	Hashtags without # symbol
`on_screen_text`	Text overlaid on video
`language`	ISO language code (e.g., `en`)
`visual_description`	2–3 sentence description: subject, setting, activity, tone
`possible_mh_relevance`	True if mental-health relevant (see Table S2)
`mental_health_keywords`	Specific MH terms found
`mental_health_confidence`	`high` / `medium` / `low`
`content_categories`	Safety flags (see Table S3)

Table S1: LLM prompt output schema. Each field is extracted from captions, hashtags, on-screen text, and visual elements visible in a single screenshot.

Signal type	Examples
Explicit diagnoses	Depression, anxiety, PTSD, bipolar disorder, eating disorders, OCD, schizophrenia
Crisis / self-harm	Suicidal ideation, self-harm, suicide prevention, crisis helplines (e.g., 988)
Treatment terms	Therapy, counseling, psychiatry, antidepressants, mental health medication
Emotional distress	"I want to die", "feeling hopeless", "panic attack", "worthless", "alone"
Recovery language	"healing", "it gets better", "day *N* of recovery", peer support
Hashtags	#depression, #mentalhealth, #anxiety, #therapy, #recovery, #sad
Visual cues	Crying / distressed subject, therapy-office setting, crisis resources displayed on screen
Excluded	*General wellness without MH framing (e.g., yoga for relaxation, nutrition only)*

Table S2: Mental-health relevance criteria used to set `possible_mh_relevance`. The classifier is instructed to prefer *inclusion* under uncertainty.

Subcategory	Operational definition
Suicide / Self-Harm	Explicit depiction of or instruction related to suicidal behavior or self-injury; graphic imagery; glorification of self-harm acts
Toxic Positivity	Dismissive "good vibes only" framing that delegitimizes clinical distress or discourages professional help-seeking
Self-Diagnosis	User-generated content diagnosing mental illness without clinical basis; misleading symptom checklists presented as definitive
Misinformation	Factually inaccurate claims about mental health causes, treatments, or medication; content contradicting evidence-based guidelines

Table S3: Potentially harmful content subcategories used in second-stage post-hoc labeling.

Appendix Section B: Extended Validation Results

The F1 scores reported in the main text (MH-relevance $= 77.42\%$; potentially harmful $= 64.52\%$; potentially supportive $= 64.00\%$) are based on a random sample of 100 videos ($\simeq 1\%$ of the corpus), independently annotated by a human rater using the same taxonomy. Per-subcategory metrics for the four harmful content types are not reported separately due to the small validation sample size.

Ablation study Multimodal inputs are necessary for real-time classification. We test whether the real-time classifier's relevance decision is robust to removing perception channels (Table S4). The full multimodal configuration attains the highest overall F1 and is the only setting that combines high recall with reasonable precision, prioritizing coverage in a safety-critical setting. Removing visual descriptions or on-screen text produces modest drops in F1 but large drops in recall, with corresponding gains in precision, indicating a systematic trade-off between coverage and conservativeness. When all visual channels are removed and the classifier relies only on captions and hashtags, F1 is roughly halved and precision collapses despite recall remaining relatively high. These patterns indicate that no single modality is sufficient for reliable real-time relevance detection: visual information and extracted on-screen text are crucial for avoiding missed mental-health content, while textual cues help refine precision. This supports our design choice to use multimodal perception for first-stage real-time classification, while reserving richer categorization for second-stage post-hoc labeling.

Appendix Section C: Operational Reliability

We assess whether the pipeline operates reliably across multi-day, multi-account collection (Table S5). Classifier latency fits within the baseline viewing window (median 5.2 s), and the second-stage post-hoc labeler is substantially faster, with typical annotations completed in under a second. All planned sessions were completed successfully, and all

Approach	What Model Sees	F1	Precision	Recall
Our Model (Full Multimodal)	Screenshots + visual_description + on_screen_text + caption + hashtags	77.42%	63.16%	100%
Vision, No Visual Desc	Screenshots + on_screen_text + caption + hashtags	73.7%	84.8%	65.1%
Vision, No On-Screen	Screenshots + visual_description + caption + hashtags	70.1%	79.4%	62.8%
No Vision [Text]	visual_description + on_screen_text + caption + hashtags	61.8%	48.8%	84%
No Vision, No On-Screen [Text]	visual_description + caption + hashtags	57.6%	44.2%	82.6%
No Vision, No Visual Desc [Text]	on_screen_text + caption + hashtags	54%	39.5%	85%
No Vision, No On-Screen, No Visual Desc [Text]	caption + hashtags	35.7%	23.3%	76.9%

Table S4: Ablation study results for binary mental-health relevance detection.

collected videos were processed by the analysis pipeline. No session-level classifier failures occurred; occasional errors triggered a conservative fallback that defaulted to skip after t_{base} ($\simeq$7 seconds) while still logging a complete trace. During collection, one account was suspended mid-study and two session cookies expired; cookies were renewed without disrupting the routine, and the suspension reduced the number of account-days but did not affect trace integrity for completed sessions. Together, these results indicate the pipeline supports sustained audits with reproducible traces and bounded operational overhead.

Metric	Value
Controller (LLM) latency	
Median (p50)	5.20 s
95th percentile (p95)	6.04 s
99th percentile (p99)	6.42 s
Labeler (Vision API) latency ($n = 120$ videos, sampled)	
Median (p50)	0.48 s
95th percentile (p95)	0.77 s
99th percentile (p99)	0.92 s
Controller reliability ($n = 138$ sessions)	
Session abort rate	0.0% (0/138)
Sessions with $\geq$1 fallback	29.0% (40/138)
End-to-end completion (all 30 accounts)	
Session completion rate	100.0% (208/208)
Video analysis rate	100.0% (8727/8727)
Avg. videos per session	42.0

Table S5: Operational reliability metrics. Latencies are wall-clock. "Fallback" denotes default action (t_{base}) used when classifier output was unavailable or invalid; "failure" denotes session abort (no completed trace).

Note: $n = 138$ covers MH-Engaged and MH-Avoidant only; Passive-Observer agents extract metadata but apply fixed watch durations. $n = 208$ covers all 30 accounts.

Appendix Section D: Exposure Trajectories and Descriptive Statistics

Figure S1 shows that behavioral interaction policy induces directional shifts in both feed composition and exposure timing, confirming that interaction policy alone—independent of initial search framing—is sufficient to produce distinct exposure trajectories under matched session budgets.

Topic composition (Figure S1A). Relative to Passive-Observer, MH-Engaged shifts the feed toward mental health (+2.27 pp) and emotional distress (+2.09 pp), with corresponding reductions in lighter topics such as dance ($-$3.76 pp) and youth culture ($-$3.07 pp). MH-Avoidant shows the mirror pattern: dance (+5.63 pp) and youth culture (+3.84 pp) increase while mental health is suppressed ($-$0.91 pp), consistent with each policy's intent to reinforce or withhold reinforcement of MH-relevant content.

Sensitive-content trajectory (Figure S1B). On Day 1, all personas begin at comparable rates ($\simeq$31–36%), reflecting shared initialization from the seed search phase. From Day 2, policies diverge sharply: MH-Engaged rises steadily to $\simeq$42% by Day 7, Passive-Observer stabilizes at $\simeq$15–19%, and MH-Avoidant declines to $\simeq$8–12% and remains stable. Sensitive content does not disappear under avoidance, indicating that a single MH-related search leaves a lasting footprint even in the absence of continued engagement—consistent with the main text findings.

Time to first harmful exposure (Figure S1C). By video index 10, only 15.7% of Passive-Observer sessions and 28.6% of MH-Engaged sessions remain harm-free, compared to 32.4% of MH-Avoidant sessions. By index 20, MH-Avoidant retains 13.2% harm-free sessions versus near-zero for the other two. Across all personas, 98.5–100% of sessions encountered at least one potentially harmful video by index 40, indicating that avoidance reduces but does not eliminate exposure to potentially harmful content—mirroring the "avoidance reduces volume but not risk" finding reported in the main text.

Descriptive Statistics. Table S6 summarizes session-level exposure outcomes by policy and framing condition. Differences in potentially harmful exposure were large and consistent across framing: MH-Engaged accounts showed mean potentially harmful rates roughly twice those of MH-Avoidant accounts (22.5% vs. 10.3%), with Passive-Observer accounts intermediate (16.0%). Help-Init. framing consistently yielded lower potentially harmful rates within each behavioral condition. FYP-specific potentially harmful rates were substantially elevated for MH-Engaged accounts (42.4%) relative to Passive-Observer (15.0%) and MH-Avoidant (8.5%).

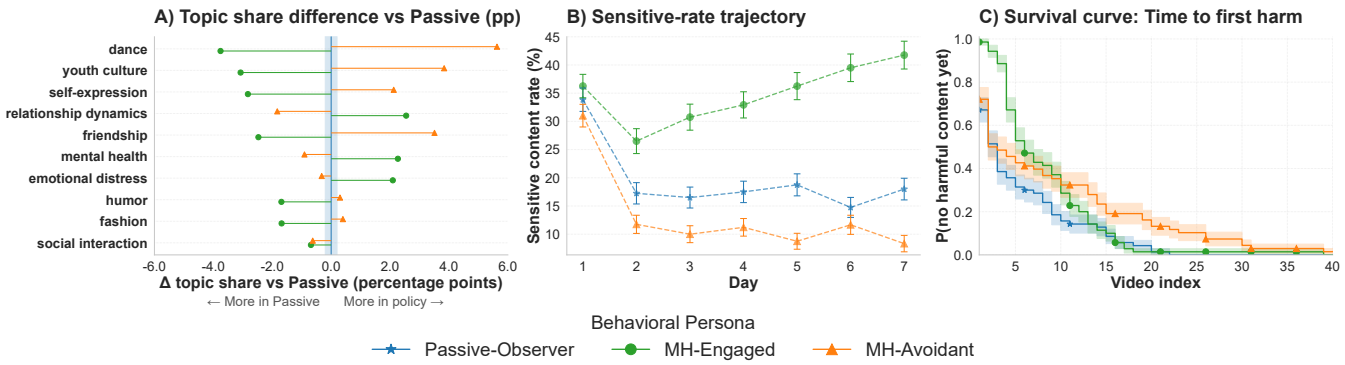

Figure S1: **Exposure trajectories across behavioral personas.** (A) Topic share differences vs Passive (percentage points). (B) Sensitive-content rate over time. (C) Survival curve: probability of no harmful content as a function of video index. (Lower curve = earlier harmful exposure)

Policy	Framing	n	Mean Videos	Potentially Harmful Rate	MH Rate	FYP Potentially Harmful Rate
Passive-Observer	Distress-Init.	35	41.8	.189 (.084)	.202 (.102)	.163 (.106)
	Help-Init.	35	41.6	.130 (.067)	.163 (.079)	.136 (.098)
MH-Engaged	Distress-Init.	35	42.2	.272 (.089)	.453 (.162)	.419 (.213)
	Help-Init.	35	41.9	.178 (.093)	.454 (.178)	.429 (.231)
MH-Avoidant	Distress-Init.	35	42.2	.117 (.065)	.126 (.064)	.093 (.046)
	Help-Init.	33	42.0	.088 (.047)	.109 (.053)	.077 (.043)
Overall	—	208	41.9	.163 (.096)	.252	.224

Table S6: Descriptive Statistics by Policy and Framing Condition

Note. Values are means with SDs in parentheses. Potentially Harmful Rate = proportion of all session videos (including search phase) labeled potentially harmful. FYP Potentially Harmful Rate = proportion of FYP-only videos labeled potentially harmful; this is the measure reported in the main text and figures. MH Rate = proportion of all session videos labeled MH-relevant.