Authorship-Resolved Detection of Suicidal Ideation in Smartphone Screenshots: Which Expressions Track Momentary Self-Report and Clinical Events

Wenpei Shao*, Brooke Ammerman*†, and Ross Jacobucci*
*Center for Healthy Minds, University of Wisconsin–Madison, Madison, WI, USA
†Department of Psychology, University of Wisconsin–Madison, Madison, WI, USA
wshao33@wisc.edu, baammerman@wisc.edu, jacobucci@wisc.edu

***Abstract*—Suicide-related language on a phone screen can be written by its owner, received from another person, or encountered as news, fiction, or an idiom. Repeated screenshots can also turn one expression into many observations. These distinctions determine what screen content measures. We developed a layer that assigns authorship from interface geometry, merges repeated captures into episodes, and distinguishes the owner's active, passive, and historical ideation from other content. In a cohort of 82 adults with recent suicidal thoughts or behaviour, 7.5 million screenshots yielded 2,077 candidate episodes, including 275 high-confidence episodes. We evaluated the resulting categories against momentary self-report and described their relation to nine confirmed clinical events. In a within-person, clock-hour-matched comparison, EMA ideation scores within 12 hours of own-active episodes were 4.5 points higher on a 0–16 scale than at referent prompts (8 participants, 33 exposed prompts; $d = 0.86$, two-sided exact sign-flip $p = .031$). Passive and colloquial categories showed no statistically detectable differences. Event-day-inclusive windows around two hospitalizations contained more detected own-ideation candidates than most safety-plan windows. Limited independent label validation and the small effective sample make these findings exploratory. The contribution is an authorship- and episode-resolved representation of screen content, with initial evidence that distinguishing expressions matters for their interpretation against self-report.**

***Index Terms*—suicidal ideation, screenomics, authorship, measurement, ecological momentary assessment, passive sensing**

I. Introduction

Suicide-related screen language may be the owner's message, a friend's disclosure, public discussion, fiction, or an idiom. Pooling these observations mixes expression with exposure. Successive screenshots can also count one persistent message repeatedly. Relating screen content to a person's state therefore requires establishing whose expression it is, what it expresses, and how many distinct episodes it represents.

These are practical measurement problems in screenomics, where screenshots are captured every few seconds during phone use [1], [2]. In a cohort of adults with recent suicidal thoughts or behaviour, the stream includes private communication alongside entertainment, public content, and the study's own survey. Prior studies of this cohort show that screen content relates to self-reported ideation. A dictionary study examined suicide-related words but could not distinguish words authored by participants from words they received or read [3]. A vision-language study reported a temporal-holdout AUC of 0.83 using screenshots in the two hours before an EMA prompt, but did not generalize to new participants [4]. Its window-level prediction did not identify which expression, or whose expression, accounted for the signal.

Our approach makes the expression episode a defined intermediate variable between screenshots and self-report. Geometry assigns text lines to sent, received, compose, or suggestion regions. Repeated captures of a conversation are merged. A language-model reader then distinguishes the participant's own active, passive, or historical ideation from received ideation, public discussion, colloquial language, and other content. Each operation resolves a different ambiguity; together they make it possible to ask what a detected episode corresponds to outside the screen.

We evaluate that correspondence using momentary EMA reports from the same participants. Ideation can fluctuate within hours [5], so the comparison is within person and matched on clock hour. We also describe screen activity around nine confirmed clinical events to illustrate the representation's coverage.

Our contribution combines authorship-resolved categories and episode aggregation with exploratory external evaluation. Own-active episodes were associated with higher subsequent EMA ideation in eight participants, motivating evaluation of defined expression categories.

II. Related Work

Public social-media research detects suicidal ideation from posts; the CLPsych task also addresses evidence extraction with large language models [6]. Screens add an attribution problem because the participant need not have written the observed material. Keyboard logging resolves authorship for typed text but does not capture incoming messages. In adolescent keyboard data, only a minority of suicide-related language represented first-person current ideation, while about a fifth was joke or hyperbole [7]. Semantic context matters even when authorship is known. Case studies of five adolescents also compared computational language measures with clinical readings before hospitalization, identifying contextual information that automated measures missed [8].

Corresponding author: Wenpei Shao (wshao33@wisc.edu).

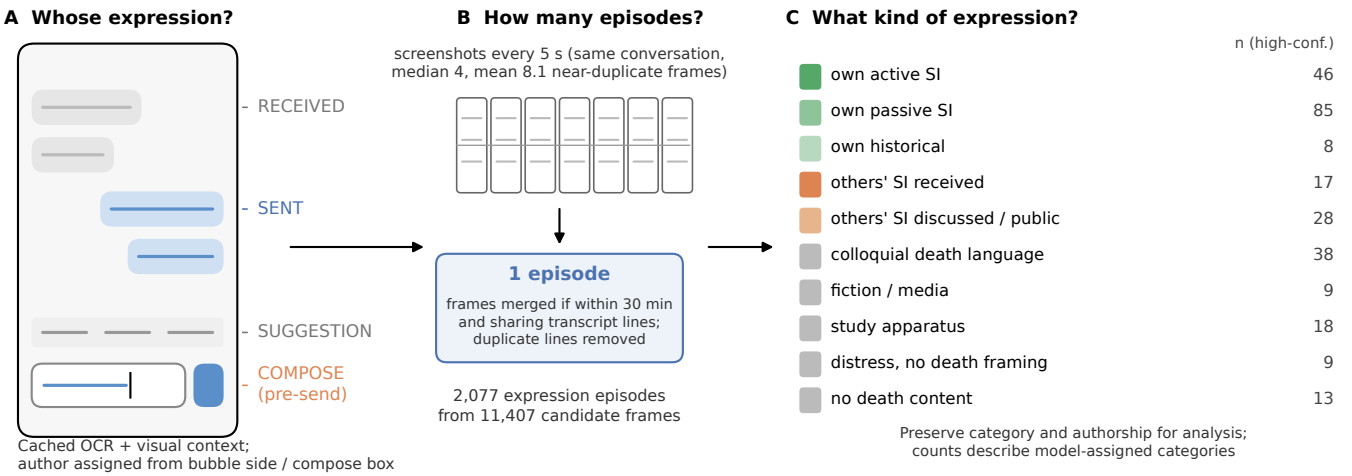

Fig. 1. Three measurement decisions between screenshots and research variables. (A) Interface geometry assigns text to sent, received, compose, or suggestion regions. (B) Repeated captures are merged into an episode; the frame-count summary refers to high-confidence own-ideation episodes. (C) The reader distinguishes categories with different meanings for the phone owner. Counts describe the high-confidence set; four crisis-service-contact episodes are omitted. Schematic only; no participant content or measured keyword-baseline comparison is shown.

A systematic review reports limited predictive value for passive behavioural signals such as location and device use [9]. Cohort analyses examine total screen time [10], nighttime engagement [11], word counts [3], and screenshot-window prediction [4]. An earlier comparative case analysis paired screen text and EMA before two psychiatric hospitalizations [12]. A related preprint trained participant-specific language-model agents on authored text and studied their simulated social interactions [13]; here we classify observed expression episodes rather than generated dialogue. These studies establish complementary analysis scales, not independent replications of the present labels. We focus on defining the observation being linked to self-report: a particular expression, attributed to its author and counted once per episode. This is also a prerequisite for studying content-triggered interventions [14].

III. Data and Detection Layer

A. Cohort and external measures

The cohort comprises 82 adults with past-month suicidal thoughts or behaviour who owned Android phones. Screenshots were captured every 5 s during phone use for five weeks: 7,522,839 screens, with a median of 85,795 per participant [4]. Six signal-contingent EMA prompts per day were delivered over 28 days [12]. The parent study had institutional review board approval and participants consented to screenshot capture. Inference on participant content ran in a restricted computing environment; this paper contains no participant text or screenshots.

The EMA ideation measure comprises two passive and two active items asking about *this moment*. Responses on 1–5 scales are rescaled and summed to $s_i \in [0, 16]$; each subscale ranges from 0 to 8. Planning items ask about the interval since the last prompt, and a separate item asks whether an attempt was made. After deduplicating repeated prompt records in the export, there are 12,914 prompts and 7,952 answered prompts from 79 participants. The clinical roster contains seven safety plans and two psychiatric hospitalizations in seven participants, with dates recorded at day resolution. Safety plans followed study outreach triggered by high active-ideation EMA reports [4]; they therefore are not an independent outcome of the survey process.

B. From screen text to attributed expression

The screen-content substrate combines EasyOCR[1] text with cached perception records from Qwen3-VL-2B-Instruct [15]. The vision model records interface and scene context; perception is separate from the pre-existing per-screen risk score used for candidate selection. The detection layer operates on 11,407 processed candidate screens (0.15% of the captured stream) selected through the risk-gated pipeline. The risk score is model-derived, not a lexical-hit indicator, and the processed subset does not establish exhaustive coverage of the captured stream.

A geometry pass assigns OCR lines to SENT, RECEIVED, COMPOSE, or SUGGESTION, using bubble position, colour, the compose region, and keyboard layout (Fig. 1A). A sent line and a received line can contain identical words but refer to different people. Compose text captures an expression before sending; suggestion text is separately marked because an interface suggestion is not equivalent to a sent message. Authorship is an inference from the interface rather than a verified account identity. On 49 hand-scored lines, 48 assignments were correct (Wilson 95% CI 0.89–1.00). This small check is not a representative estimate across all apps or layouts. Partly occluded bubbles and garbled OCR are known failure modes.

[1] https://github.com/JaidedAI/EasyOCR

TABLE I
FROM CAPTURED SCREENS TO ANALYSIS SETS. COUNTS DESCRIBE MODEL-ASSIGNED CATEGORIES.

Stage	Count
Screenshots captured (82 participants)	7,522,839
Candidate screens processed by the layer	11,407
Episodes after merging	2,077
High-confidence set	275
Own active / passive / historical	46 / 85 / 8
Others' ideation received / discussed	17 / 28
Colloquial / fiction / apparatus	38 / 9 / 18
Distress / no death / crisis contact	9 / 13 / 4
Review queue	1,802
Own-ideation candidates	348
Answered EMA prompts (79 participants)	7,952

TABLE II
EVIDENCE AVAILABLE AT DIFFERENT LEVELS OF THE PIPELINE. EACH CHECK ADDRESSES A DIFFERENT QUESTION.

Level	Evidence	Interpretation boundary
Line authorship	48/49 hand-scored assignments	Small interface-level check; not an episode-label test.
Episode category	15/17 high-confidence episodes in 44 adjudicated alerts	Development-reused, single-rater sample; no independent precision estimate.
EMA association	Active: 8 persons, 33 exposed prompts	Within-person association conditional on detected labels and answered prompts.
Clinical context	9 events in 7 persons	Descriptive, day-resolution windows; not prospective event sensitivity.

C. From repeated frames to classified episodes

Screens from the same conversation are merged when they occur within 30 minutes and share transcript lines, with line-level deduplication inside the episode. The 11,407 screens form 2,077 episodes. Among the 139 high-confidence own-ideation episodes, the mean number of frames is 8.1 (median 4; maximum 122). A frame count would therefore partly measure how long content stayed visible rather than how often it was expressed. An episode is an operational grouping of observed content, not a claim that the underlying psychological state begins or ends at those boundaries.

An open-weight 30B-class instruction model reads the author-labelled transcript and assigns an episode category. The categories distinguish own active ideation, own passive ideation, own historical ideation or attempt, others' ideation received in a private exchange, others' ideation or death discussed or viewed publicly, colloquial death language, fiction or media, study apparatus, distress without death framing, no death content, and crisis-service contact. A deterministic idiom gate demotes clean idioms. The active/passive distinction follows the constructs represented in the EMA items and C-SSRS [16]: active expressions concern ending one's life or thinking about doing so; passive expressions concern being dead or not waking up without an active component.

The *high-confidence set* comprises 275 episodes selected using death-language cues and vision-based death-content judgements, with reader-based filtering of fiction and third-party health content. This is a pipeline selection rule, not independent confirmation that every retained category is correct. It contains 139 own-ideation episodes: 46 active, 85 passive, and 8 historical (Table I). The other 1,802 episodes form a review queue, including 348 own-ideation candidates. We use high-confidence categories for the EMA analysis; review-queue counts are reported only as model-generated candidates in the clinical descriptions.

D. What has been validated

One researcher adjudicated 44 alert-selected episodes. Of the 17 belonging to the high-confidence set, 15 were categorized correctly. The disagreements involved a metaphor read as passive ideation and a crisis-service contact judged by the adjudicator to express own ideation. The sample was consulted during development and is neither random nor independent. It cannot provide unbiased category-specific precision or end-to-end recall. There is no independent clinician-labelled evaluation set in the present study.

Table II separates the questions addressed by each check. Line attribution evaluates a component; episode annotation evaluates category assignment; EMA assesses external correspondence. Millions of captured frames describe the setting's scale, whereas the participants contributing to each contrast determine its empirical support.

E. Retrospective alert stream

For clinical descriptions, we replay rules for crisis contact and strong compose/sent ideation phrases. Crisis-contact and strong-phrase bypasses alert directly; other rule alerts are verified by a Qwen3-4B classifier [17] with a LoRA [18] sequence head distilled from reader labels. In the development-reused adjudication set, retained-queue precision was 15/22 (0.68; Wilson 95% CI 0.47–0.84), versus 18/44 (0.41; 0.28–0.56) for the original queue. Replay burden was 0.063 alerts per participant-day over 1,950 EMA-window days. Human-confirmed behavioural alerts are counted separately.

IV. VALIDATION DESIGN

A. Within-person EMA comparison

The expression episode defines exposure; the EMA comparison uses answered prompts and inference across participant-level differences. A prompt is exposed if the most recent high-confidence episode of that category began within the preceding 12 h. One episode may expose several prompts. We examine own active ideation (8 participants, 33 exposed prompts), own passive ideation (18, 100), and colloquial death language (9, 52).

The analysis uses a case-crossover comparison [19]: referents are the same participant's answered prompts on other dates, within two clock hours, and with no own-ideation episode of any confidence in the preceding 48 h. The implemented clock-hour match uses the absolute difference of hour-of-day values without wrapping across midnight. There are

356 case–referent matches for the 33 active-exposed prompts; referent prompts may be reused and this is not a count of unique referents. Each exposed prompt has a mean of its matched referents subtracted from its score. These differences are averaged within participant, then across participants with equal participant weight.

We report the mean difference, Cohen's d across participant differences, the number of positive differences, and a two-sided exact sign-flip test. We enumerate all 2^N participant sign assignments and count those with an absolute mean at least as large as observed. The test assumes symmetry of participant differences under the null. A Wilcoxon signed-rank test provides a second summary; the leave-one-participant-out range assesses sensitivity to omitting one person and is not a confidence interval.

This is an exploratory analysis. The broader analysis workflow examined several exposure windows and category definitions; the 12 h contrast is the focal analysis reported here, not a preregistered confirmatory test. Sensitivities retain participants with at least two exposed prompts, widen matching to ± 4 h with a 24 h clean period, and remove episodes beginning within 15 minutes of an EMA prompt. The last check addresses possible survey-related contamination. Category and subscale tests are unadjusted for multiplicity. We do not test the difference between category effects, and significance for one category but not another is not evidence of such a difference.

B. Clinical context and attempt-item reports

For each event, we describe the 168-hour window ending at 23:59 on its recorded date. It includes the event day, so same-day observations may precede or follow the event. We report own-ideation candidates of any confidence, high-confidence active/passive episodes, distress candidates, retained replay alerts, and the latest EMA by the endpoint and its maximum in the window. These nine instances are descriptive, without inferred warning lead times.

The attempt item was endorsed on 28 answered prompts by 18 participants. Fourteen endorsements from nine participants co-occurred with nonzero ideation or planning; the other 14 had both scores equal to zero. An exploratory consistency-filtered analysis retained the former subset. This unvalidated filter does not establish that excluded responses were errors, or that retained endorsements were confirmed, distinct attempts.

V. Results

A. The representation separates different kinds of content

Of the 275 high-confidence episodes, 139 are classified as the participant's own ideation, including eight historical episodes. Others' ideation received or discussed accounts for 45 episodes, colloquial language for 38, and fiction, study apparatus, distress, no death content, or crisis contact for the remaining 53. Thus, even within this selected set, an undifferentiated count would pool several kinds of observations. These are category distributions produced by the pipeline, not estimates of the performance of a keyword detector.

TABLE III
EXPLORATORY EMA CONTRASTS. N: PARTICIPANTS; m: EXPOSED PROMPTS; DIFF.: MEAN PARTICIPANT-LEVEL EXPOSED-MINUS-REFERENT DIFFERENCE; d: STANDARDIZED DIFFERENCE; POS.: POSITIVE PARTICIPANT DIFFERENCES; p: TWO-SIDED EXACT SIGN-FLIP, UNADJUSTED.

Episode set	Outcome	N	m	Diff.	d	pos.	p
Own active	si	8	33	+4.53	.86	6/8	.031
	active	8	33	+2.22	.84	6/8	.031
	passive	8	33	+2.31	.87	6/8	.047
	Pr(si ≥ 4)	8	33	+.35	1.01	6/8	.031
≥ 2 exposures	si	6	31	+3.85	.92	5/6	.063
Wider match	si	9	34	+4.51	.94	7/9	.020
Not EMA-adj.	si	7	24	+4.80	.85	5/7	.063
Own passive	si	18	100	−.33	−.29	7/18	.233
Colloquial	si	9	52	+.62	.37	5/9	.293

The distinction between observation scales is also substantial: the 139 own-ideation episodes contain an average of 8.1 frames each. A long-visible message can contribute many frames without being a new expression. Among 106 high-confidence own-ideation episodes within participants' EMA windows, the median distance to the nearest answered prompt is 2.5 h; 16% have no answered prompt in the following 24 h. The EMA comparison consequently evaluates only episodes and prompts for which a temporal link is observable, not every expression detected in the stream.

B. Own-active episodes and subsequent EMA ideation

At answered prompts within 12 h of an own-active episode, ideation is on average 4.53 points higher than at clock-hour-matched referents (8 participants, 33 exposed prompts; $d = 0.86$; two-sided exact sign-flip $p = .031$; Wilcoxon $p = .039$). Six of eight participants have positive differences. Active and passive EMA subscales are both higher, by 2.22 and 2.31 points respectively (Table III). The association therefore is not confined to the active subscale's wording.

The participants contribute between one and ten exposed prompts each. Equal participant weights include three participants with differences of at least eight points and only one to three exposures. Figure 2 shows this heterogeneity. Omitting any one participant leaves a positive mean difference, ranging from 3.26 to 5.27 points. The direction is stable to a single omission.

The sensitivity estimates remain positive and similar in magnitude: 3.85 points when requiring at least two exposed prompts, 4.51 with wider matching, and 4.80 after excluding EMA-adjacent episodes. Their two-sided exact p values are .063, .020, and .063. Thus, the direction and magnitude persist across these specifications, while two sensitivities do not meet the .05 threshold. The matched estimates remain conditional on the selected risk gate, episode labels, observed prompt responses, and the exposure window.

C. Other expression categories

Own-passive episodes have a mean difference of -0.33 points across 18 participants and 100 exposed prompts ($d =$

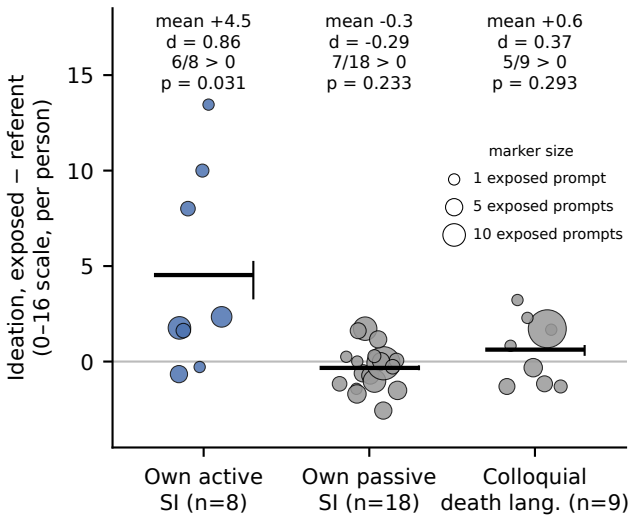

Fig. 2. Participant-level EMA differences for each detected category. Marker size indicates the number of exposed prompts. Horizontal bars mark category means; vertical ticks at their right ends show leave-one-participant-out ranges, not confidence intervals. Displayed p values are two-sided exact sign-flip tests. The categories involve different, potentially overlapping participant sets.

-0.29, two-sided exact $p = .233$). Colloquial episodes have a mean difference of $+0.62$ across nine participants and 52 prompts ($d = 0.37$, $p = .293$). Neither contrast detects a difference from zero; neither establishes equivalence to zero or a difference from the active-category effect.

Participants expressing passive ideation had average EMA scores similar to those expressing active ideation (3.8 versus 4.3), and passive-exposed prompts averaged 3.8. A high baseline alone is not an evident explanation. Expression timescale, category-specific label error, and alignment with the EMA construct remain possible explanations.

Own-ideation episodes were adjacent to an EMA prompt within 15 minutes in 21% of cases versus 27% for other episodes. Together with the positive estimate after excluding adjacent episodes, this reduces concern about immediate survey contamination, without ruling out other study-related or classification effects.

D. Clinical windows illustrate coverage and omissions

Table IV describes event-day-inclusive windows around the nine clinical events. The two hospitalization windows contain 12 and 23 own-ideation candidates of any confidence and three and eight retained alerts, respectively. The second also contains nine high-confidence active episodes. Most safety-plan windows contain fewer candidates: six have zero to three own-ideation candidates and no retained alert; the remaining window has six candidates and two alerts. Because lower-confidence candidates are included, these counts describe pipeline activity rather than verified expression frequencies.

High EMA ideation can coexist with little detected own-ideation content. P5 has an EMA maximum of 16 in the safety-plan window, with no own-ideation candidate and two distress candidates. This illustrates incomplete correspondence between the measures. Event-day ambiguity and the EMA-triggered safety-plan protocol limit interpretation of the hospitalization/safety-plan contrast.

TABLE IV
SEVEN-DAY WINDOWS ENDING AT 23:59 ON EACH EVENT DATE, INCLUDING THE EVENT DAY. OWN: OWN-IDEATION CANDIDATES OF ANY CONFIDENCE; A/P: HIGH-CONFIDENCE ACTIVE/PASSIVE; D: DISTRESS CANDIDATES; ALERTS: RETAINED REPLAY ALERTS; EMA: LATEST BY THE ENDPOINT / MAXIMUM IN THE WINDOW.

	Event	Own	A/P	D	Alerts	EMA
P1	Safety plan	2	0/0	0	0	8/8
P2	Safety plan	0	0/0	0	0	0/5
P3	Safety plan	6	2/2	0	2	6/15
P3	Hospitalization	12	1/3	2	3	6/13
P4	Safety plan	3	0/1	1	0	10/12
P4	Hospitalization	23	9/5	4	8	16/16
P5	Safety plan	0	0/0	2	0	16/16
P6	Safety plan	0	0/0	0	0	14/14
P7	Safety plan	1	0/0	1	0	7/7

Of the 14 consistency-filtered attempt-item endorsements, five have an own-ideation candidate in the preceding seven days and three have a retained alert. These describe endorsed prompts, not verified attempts or detection sensitivity. The last screenshot for each roster event was captured on the event day or preceding day, but intermittent gaps within the window remain possible.

VI. Discussion

A. Defining what a screen-derived variable means

The representation distinguishes expression from exposure and an episode from repeated captures. Received disclosures, historical disclosures, passive wishes, active expressions, and idioms may all matter, but their shared vocabulary does not make them equivalent observations. Preserving these distinctions specifies what downstream analyses link to an external criterion.

The EMA result offers initial support for that approach. Detected own-active episodes are associated with higher subsequent self-reported ideation within participant, including both active and passive subscales. The personal baseline and clock-hour match make this a more specific question than whether people with more suicide-related words have higher average ideation. However, it remains an association over a 12-hour observation window. Expression could reflect an already elevated state, and a later report does not establish the onset, duration, or direction of change of that state.

The passive and colloquial results are useful because they prevent the positive active-category estimate from being generalized automatically to all detected content. They do not demonstrate that passive expressions lack clinical meaning. A passive expression may correspond to a different timescale or outcome, and the classifier may separate the categories imperfectly. The variables should remain distinct while their meanings are investigated.

B. Implications for downstream studies

Retaining authorship, category, confidence, episode boundaries, and survey proximity lets downstream studies examine whether associations depend on semantic category, received content, or persistent messages.

Clinical descriptions show why coverage matters: screens observe material visible during phone use, while EMA elicits a scheduled report. High EMA scores can coexist with few detected expressions. Intervention research would additionally require independent alert validation, prospective event timing, and evaluation of burden and missed events; the present retrospective analysis supplies a measurement layer for that work.

C. Limitations and next validation steps

The focal EMA contrast rests on eight participants, several with few exposed prompts, despite the much larger captured stream. Per-person estimates and sensitivity analyses make that concentration visible but cannot substitute for replication. The cohort is a single Android sample followed for five weeks; the analysis is conditional on answered EMA prompts and a risk-gated subset of screens. Capture occurs only during phone use, and coverage outside the processed gate has not been established by a representative end-to-end recall evaluation.

Episode labels are model-generated. The small single-researcher check was reused during development, and neither the 48/49 line check nor the EMA association replaces independent episode annotation. Label error could attenuate or distort category associations; its direction is unknown. The overlapping exposure windows and reused referents also mean the prompt counts should not be interpreted as independent replications. The participant-level tests address clustering at the summary level, but their assumptions and the small sample limit inferential certainty. The exploratory choice of windows, multiple unadjusted contrasts, and non-circular clock-hour matching are further reasons to treat the results as provisional.

The next priority is independent, blinded annotation stratified by category and participant, separating authorship errors from semantic errors and sampling outside the high-confidence set. Participant-held-out evaluation and replication would then test whether the observed association extends beyond this cohort.

VII. Conclusion

Retaining authorship, semantic category, and episode boundaries defines distinct screen-derived research variables. In this cohort, detected own-active expressions showed an exploratory within-person association with higher subsequent EMA ideation. Other category contrasts and clinical descriptions identify further validation needs; they establish neither absence of signal nor clinical warning performance.

References

[1] B. Reeves, T. Robinson, and N. Ram, "Time for the human screenome project," *Nature*, vol. 577, no. 7790, pp. 314–317, 2020.

[2] B. Reeves, N. Ram, T. N. Robinson *et al.*, "Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them," *Human–Computer Interaction*, vol. 36, no. 2, pp. 150–201, 2021.

[3] B. A. Ammerman, E. M. Kleiman, C. O'Brien *et al.*, "Smartphone-based text obtained via passive sensing as it relates to direct suicide risk assessment," *Psychological Medicine*, vol. 55, p. e144, 2025.

[4] R. Jacobucci, W. Shao, V. Kobrinsky, and B. Ammerman, "Predicting momentary suicidal ideation from smartphone screenshots using vision-language models: Prospective machine learning study," *JMIR Mental Health*, vol. 13, p. e90581, 2026.

[5] E. M. Kleiman, B. J. Turner, S. Fedor, E. E. Beale, J. C. Huffman, and M. K. Nock, "Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies," *Journal of Abnormal Psychology*, vol. 126, no. 6, pp. 726–738, 2017.

[6] J. Chim, A. Tsakalidis, D. Gkoumas *et al.*, "Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts," in *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. Association for Computational Linguistics, 2024, pp. 177–190.

[7] P. A. Bloom, I. N. Treves, D. Pagliaccio *et al.*, "Identifying suicide-related language in smartphone keyboard entries among high-risk adolescents," *npj Digital Medicine*, 2026, advance online publication.

[8] I. N. Treves, P. A. Bloom, S. Salem *et al.*, "In their own words: Case studies of adolescent smartphone language preceding suicide-related hospitalizations," *NPP—Digital Psychiatry and Neuroscience*, vol. 4, p. 5, 2026.

[9] R. Büscher, T. Winkler, J. Mocellin *et al.*, "A systematic review on passive sensing for the prediction of suicidal thoughts and behaviors," *npj Mental Health Research*, vol. 3, p. 42, 2024.

[10] R. Jacobucci, M. Blacutt, N. Ram, and B. A. Ammerman, "Smartphone screen time and suicide risk in daily life captured through high-resolution screenshot data," *npj Digital Medicine*, vol. 8, p. 321, 2025.

[11] R. Jacobucci, S. G. Jones, M. Blacutt, and B. A. Ammerman, "Passive vs active nighttime smartphone use as markers of next-day suicide risk," *JAMA Network Open*, vol. 8, no. 11, p. e2542675, 2025.

[12] R. Jacobucci, B. Ammerman, and N. Ram, "Examining passively collected smartphone-based data in the days prior to psychiatric hospitalization for a suicidal crisis: Comparative case analysis," *JMIR Formative Research*, vol. 8, p. e55999, 2024.

[13] W. Shao, B. Ammerman, and R. Jacobucci, "Surfacing suicidal risk through simulated social interaction: Per-person language model agents as communicative stress tests," *medRxiv*, p. 2026.06.04.26354928, 2026, preprint.

[14] I. Nahum-Shani, S. N. Smith, B. J. Spring *et al.*, "Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support," *Annals of Behavioral Medicine*, vol. 52, no. 6, pp. 446–462, 2018.

[15] S. Bai, Y. Cai, R. Chen *et al.*, "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[16] K. Posner, G. K. Brown, B. Stanley *et al.*, "The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults," *American Journal of Psychiatry*, vol. 168, no. 12, pp. 1266–1277, 2011.

[17] A. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[18] E. J. Hu, Y. Shen, P. Wallis *et al.*, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[19] M. Maclure, "The case-crossover design: A method for studying transient effects on the risk of acute events," *American Journal of Epidemiology*, vol. 133, no. 2, pp. 144–153, 1991.