

1. Key Findings - Recommended Variables. By applying a mixed-methods approach to NVDRS Narrative data, we found evidence of a diverse set of events and experiences related to decedent engagement in online spaces. Aligned with our theoretical framing (described in Background), we propose the following five variables be added to NVDRS: “Private Sharing”, “Conflict”, “Victimization”, “Withdrawal”, and “Harmful Exposure”. A description of the codes is provided in Table 1, but briefly; private sharing refers to the decedent sharing internalized thoughts, feelings or other emotions via social media, not including suicide disclosure. This could include anger, stress, or sadness, all of which can be precursors to suicide. Conflict includes digital interactions with others, such as a dating partner, which can exacerbate stress and risk of suicide. Victimization refers to social interactions whereby the decedent is perpetrated against, for example cyberbullying or the sharing of sexually explicit material of the decedent without their permission; these types of interactions can exacerbate suicidality. Withdrawal from social media includes shutting down accounts, or a caregiver restricting access to accounts or devices; withdrawal is a common precursor of suicidality, and as social media spaces are important and common for adolescents, withdrawal on social media may be an important indicator of suicidality. Finally, we identified that individuals were seeing harmful content (Harmful Exposure) in online spaces, including explicit images, violence, or other portrayals of harms being inflicted on another person or thing.

2. Background. By age 17, 95% of contemporary youth have access to a smartphone, and adolescents spend on average 4.8 hours a day on social media applications (apps).¹ Research links higher use of social media, longer screen times, and harmful online activities (i.e., cyberbullying, coercive sexting) with increased mental health challenges and youth suicide ideation.²⁻⁷ Given the ubiquitous use of online spaces among youth and the complexity of online interactions, it is unclear which specific mechanisms and online behaviors are linked to youth suicide deaths. However, current research is limited by lack of contextual information on social media engagement, and there is a need to identify pathways and mechanisms that increase risk.^{8,9} Recent rise in youth suicide¹⁰ highlights the urgent need to understand how online experiences contribute to this public health issue.

While NVDRS contains a ‘Disclosed to Social Media’ variable, it does not otherwise characterize how youth may engage in online spaces prior to suicide. *We draw on two key theories that help characterize heterogeneity of non-normative online behavior in the context of suicide and identify relevant themes.* First, Durkheim’s theory of Suicide¹¹ provides a framework for understanding why violation of established social norms is a useful heuristic. Durkheim suggests that a balance between the social norms of integration and regulation acts as a glue for any social group. An imbalance -- whether excess or lack of regulation or integration -- can foster ‘deviant’ or non-normative behavior and increase the risk of suicide among group members. Second, the Integrated Motivational-Volitional (IMV) theory¹² helps characterize the online behaviors that may signal a person’s readiness to transition from suicidal ideation to active planning and suicide. The IMV outlines progression toward suicide using 3 risk phases: pre-motivational (a trigger for suicidality), motivational (experiencing suicidal ideation in response to trigger), and volitional (suicidal action via behavioral normalization). Through various mechanisms, online spaces can exacerbate feelings of isolation, low self-esteem, and entrapment leading to negative emotional states, evident in instances of sharing private or intimate content, and sharing content that encourages suicide, discusses means of injury, or exposes an individual to violent content. Variable creation was guided using these theoretical framings (Table 1).

3-4: Methods, Approach, and Rationale. The goals of this project are to use mixed-methods to develop five novel variables; specifically, we 1) use open-coding of NVDRS narratives to identify key themes related to decedents’ non-normative behaviors online, and 2) use large language models (LLMs) to classify each narrative into these themes. First, we performed qualitative thematic analysis to define a set of five variables describing


Code	Definition	Example	IMV	Integration/Regulation
Private sharing	Decedent shared online their own private thoughts or emotional states/reactions that exposed their vulnerability (e.g., identities, fears, negative emotional states, or self-harm).		Pre-Motivational	Hi / Lo
Conflict	Interpersonal argument or other conflict entered into because of online content or expressed online.		Pre-Motivational	Hi / Lo
Victimization	Narrative explicitly says the decedent was harmed (or emotionally affected) in an online space. Harm may include cyberbullying, harassment, unauthorized sharing of information, or someone doing some other unwanted thing to them.		Pre-Motivational	Lo / Lo
Withdrawal	Refers to evidence of abrupt or complete reduction of all online activities including as a disciplinary act from a parent.		Motivational moderators	Lo / Hi
Harmful Exposure	Narrative explicitly says the decedent consumes content showing harm (physical, verbal, emotional, etc.) towards (real or fictional) others in an online space.		Volitional moderators	Lo / Lo

Table 1. Does not include private sharing online suicide note or disclosure. Does not apply if decedent is self-harming.

non-normative behaviors in online spaces. To obtain a sub-sample with relatively high density of relevant cases, a set of 51 keyphrases and heuristic inclusion/exclusion rules were developed to identify a subset of 1,279 narratives that are likely to reference the decedent’s online behavior. To identify these keyphrases, we started with a set of 15 hand-collected keyphrases (e.g., social media, phone, video) and augmented these using techniques from natural language processing like log-odds ratios of word frequencies and word embeddings to identify words occurring in similar narratives. These narratives contain words or phrases often used to describe online activity (e.g., “online” or “post”) (Table 2). Eight authors conducted an initial exploratory review of 50 each, for a total of 400 annotated narratives. Each author identified 17 - 22 potentially relevant cases in their subset and summarized the engagement in online spaces reflected in the narrative. Then, the authors collaboratively developed a set of 12 sub-themes (5 of which we recommend should be new variables, the other 5 being contextualizing pieces of information such as a disclosure of suicidality on social media, and law enforcement explicitly reviewed decedent’s social media). The 5 proposed themes capture the nuances of the behaviors described in the narrative (e.g., “decedent views content showing harm (physical, verbal, emotional, etc.) towards (real or fictional) others in an online space,” [Harmful Exposure]). These themes capture qualitatively and quantitatively distinct attributes about online behavior than what is already in the NVDRS variables. Notably, **96% of cases identified by the keyphrases were not already identified by the existing ‘Disclosed to Social Media’ variable**. To complete this step, we coded an additional set of 300 narratives to refine the themes.

Types of Phrases	# Cases	SM Disclosure = T	SM Disclosure = F
7 Web Phrases (<i>online, cyber, web, etc.</i>)	145	4 (3%)	141 (97%)
12 Social Media Phrases (<i>facebook, etc.</i>)	136	9 (6%)	127 (94%)
13 Message Phrases (<i>post, chat</i>)	821	46 (6%)	775 (94%)
19 Other Phrases (<i>game, stream, etc.</i>)	475	12 (3%)	463 (97%)
All Phrases	1279	50 (4%)	1229 (96%)

Table 2. Cases for annotation in thematic analysis, identified by 51 key phrases for online spaces

In the second step of this project, we explored the potential of LLMs to identify complex themes pertaining to decedent engagement in online spaces from the NVDRS narratives. While prior literature fine tuned LLMs to detect themes in the narratives,¹³ this approach requires large amounts of human-labeled data for training and validation. Since LLMs are trained on large amounts of natural language data, they have achieved high performance in some classification tasks without additional finetuning (i.e., zero-shot classification), expanding access to these methods.¹⁴ To more systematically evaluate the zero-shot performance of LLMs at this task, we manually annotated 472 unique narratives; 388 of these were singly annotated and 84 were multiply-annotated (by 2-4 annotators each) for inter annotator agreement, so there were 642 annotations in total.

We then adopted an iterative prompting approach to identifying and classifying decedent online engagement in the narratives. Our main results were calculated using Meta’s open-source LLM Llama3.1-8B-Instruct.¹⁵ We also tested the performance of our prompts using Mistral-7B-Instruct-v0.3 which had worse performance. We tried many different prompting strategies in developing our final approach, and we developed this approach to account for numerous challenges faced using a smaller open-sourced LLM (i.e., Llama3.1-8B-Instruct).

1. First, we prompted the LLM to break the narratives into individual sentences to account for the model’s difficulty in processing the full narrative. When prompted using a full narrative, the model had much lower recall and similar precision, as the model often did not recognize relevant sentences that were embedded in the narrative. This likely occurs because smaller LLMs have smaller context windows.
2. Second, we noticed that the LLM faced difficulty distinguishing between online behaviors of interests identified in our thematic analysis; for instance, the LLM would often include offline conflicts in the Conflict code, offline bullying in the Victimization code, and so on. Therefore, we prompted the LLM to recognize whether each sentence references online spaces utilizing the following prompt: *“Does the following sentence talk about an online space? This includes social media, web searches, messaging, chat, email, viewing or posting content, gaming, online schooling, phones, computers, or cyberbullying. This does not include texting.”*
3. Finally, the LLM was often unable to distinguish between various aspects of online engagement in these narratives; for instance, there were often aspects of suicide disclosure, even though we had excluded many instances of disclosure from how we defined these other codes (e.g., someone posting “goodbye” on Facebook was seen as withdrawal). To account for this, we craft a multiple-choice prompt for each code to create clear inclusion and exclusion criteria for each prompt (this is illustrated in Code Snippet 2 below).

5. Analytic Performance.

1,279 narratives (31.98%) were extracted based on keywords from the provided dataset of 4,000 narratives. 472 unique narratives were randomly chosen to be hand-coded; of those, 299 (63.4%) of narratives contained reference to social media. Within the 299 narratives, 11.02% contained reference to Private Sharing, 3.60% to Conflict, 4.03% to Victimization, 5.51% to Withdrawal, and 7.20% to Harmful Exposure (Table 3). To

ensure the reliability of our thematic analysis, we conducted an inter-annotator agreement (IAA) using Krippendorff's alpha and pairwise agreement. IAA ranged from 0.40 (Private Sharing) to 0.77 (Conflict), with 4 out of 5 codes having Krippendorff's alpha >0.65. One of the challenges for Private Sharing was differentiating sharing that was more aligned with disclosure versus sharing that was emotional in nature but not disclosure. In these cases, it was challenging to determine if the sentence indicated Private Sharing or was suicidal disclosure. To evaluate the zero-shot performance of the LLMs in identifying our five variables, we used precision, recall, and F1 statistics. Precision of models ranged from 0.31 (Private Sharing) to 0.90 (Harmful Exposure) and 3 out of 5 of the variables achieved precision of ≥ 0.65 . Precision for the variable of Conflict was on the lower end (0.39); additional prompt engineering would likely improve precision for this variable. Recall was sufficient across variables, ranging from 0.62 (Private Sharing) to 0.80 (Withdrawal). F1 statistics largely mirrored the IAA and Precision statistics, with lower F1 scores for Private Sharing (0.42) and Conflict (0.51), and though 3 out of 5 variables had F1 statistics of >0.65 .

Out of the 4,000 provided narratives, the LLM identified 804 (20.10%) as being related to social media. [The LLM prevalence is smaller than the hand-coded data, as the hand-coded data narratives first used the keyword query to elicit relevant narratives to hand code.] Within the narratives identified as containing reference to social media (N=804), 25.00% were classified as referring to Private Sharing, 7.59% were classified as containing reference to Conflict, 3.73% contained reference to Victimization, 5.10% contained reference to Withdrawal, and 6.59% contained reference to Harmful Exposure. The LLM predicted prevalence was similar to the hand-coded prevalence for Victimization, Withdrawal, and Harmful Exposure. Notably, the LLM predictions were most dissonant with hand-coded prevalences for Private Sharing and Conflict, giving further motivation for additional review and refinement for inter-rater reliability (Private Sharing) and prompt engineering (Conflict).

6. Alternative Approaches. Given the regulations of the current competition, we were unable to use larger LLMs including a HIPAA compliant LLM offered through the University of Michigan (i.e., UM-GPT, which is proprietary, thus not available to the broader public). This limited us to the size of LLM we were able to use, which were not able to process the full narratives and required that we feed individual narrative sentences through the model. The challenge with sentence-based classification is that many narratives contain one sentence noting social media-related factors, meaning the remaining sentences are not relevant to the classification. When we beta tested the larger models such as UM-GPT, we found that the LLM was able to utilize the full narrative and had strong accuracy.

7. Critical Code.

Code Snippet 1 (left below): The following code snippet shows the function used to calculate interannotator agreement of our hand-annotated sample. In order to do this, we read in the hand annotations in long-format, reshape the data to be in the format required for the Krippendorff's alpha function, and then calculate the nominal value of alpha for each of the codes in our sample.

Code	Krippendorff's Alpha	Pairwise Agreement	Prevalence n(%)	Precision	Recall	F1	LLM Predicted Prevalence
Narratives w/ Reference to SM	0.69	83.43%	299 (63.35%)	0.82	0.90	0.86	20.10%
Private sharing	0.46	85.71%	52 (11.02%)*	0.31	0.62	0.42	25.00%
Conflict	0.77	98.86%	17 (3.60%)*	0.39	0.75	0.51	7.59%
Victimization	0.66	97.14%	19 (4.03%)*	0.65	0.73	0.69	3.73%
Withdrawal	0.74	97.71%	26 (5.51%)*	0.70	0.80	0.74	5.10%
Harmful Exposure	0.76	97.71%	34 (7.20%)*	0.90	0.66	0.76	6.59%

Table 3. *The prevalence is relative to narratives with reference to social media (N=411).

```
def alpha(annotations, codes):
    code_alphas = pd.DataFrame(columns=['Code', "Krippendorff's Alpha"])

    for code in codes:
        reliability_data = annotations[['Person', 'uid', code]]
        [annotations['For IAA'] == 1].drop_duplicates(subset=['Person',
            'uid']).pivot(index='Person', columns='uid', values=code)
        reliability_data_input = [[value for value in row] for row in
            reliability_data.values]
        pd.DataFrame(reliability_data_input).to_csv('reliability.csv')

        if annotations[annotations['For IAA'] == 1][code].sum() != 0:
            alpha = krippendorff.alpha(reliability_data=reliability_data_input,
                level_of_measurement='nominal')
        else:
            if VERBOSE: -
                alpha = pd.NA

        code_alphas = code_alphas.append({'Code': code, "Krippendorff's
            Alpha": alpha, ignore_index=True})

    return code_alphas
```

```

<Text of the Prompt
template = """
[begin_of_text]<[start_header_id]>[system_header_id]>

I am a researcher studying suicide risk factors. You are a helpful AI question answering assistant, who answers all my questions.<[eot_id]>[<
In the following sentence, which of the following is true? Give only the letter with no explanation.

A. V posted on social media or messaged someone indicating they were thinking about suicide or planning to kill or hurt themselves
B. Someone took away V's access to internet, phone, computer, gaming, social media, or other devices
C. V had stopped using social media, deleted an account, or withdrew from an online account
D. None of the above

Sentence: {narr}<[eot_id]>
<[start_header_id]>[assistant_header_id]>
Answer:
"""
"""

# Create prompts for all sentences
prompt_list = [template.format(narr = x) for x in sents_headered_sentence]

# Run inference on all prompts using an open source LLM loaded from huggingface (defined previously)
codes = llm.generate(prompt_list, sampling_params=SamplingParams(max_tokens=8192, temperature=0))
print(Counter([x.outputs[0].text for x in codes]))

# Apply the withdraw code if the LLM chose B or C
codes = 'withdraw'
sents_labeled[code+'llm'] = 'n'
sents_labeled[code+'uid',code+'lm'] = [x.outputs[0].text for x in codes]
sents_labeled[code+'lm'] = sents_labeled[code+'llm'].apply(lambda x: 1 if x in ['B','C'] else 0)

# Calculate interannotator agreement at a narrative level
agg = sents_labeled.groupby(['uid','person'])[['disclosure_llm',code+'llm']].max().reset_index().\
merge(laa_narrs[['uid','code']],on='uid',how='left')
evaluate_code(agg)

```

Code Snippet 3: In this final code snippet, we calculate the confusion matrix, precision, recall, and F1 score of the LLM against the ground truth of our hand annotations for each code.

Additionally, we printed all sentences that were examples of false positives and false negatives to help us engineer better prompts. Prompts were engineered from a subset of the hand-annotated set (the equivalent of a “dev” set) and test performance was reported from the avoid overfitting of the prompt to the idiosyncrasies of the hand annotated

```
for i in sents_labeled[(sents_labeled[code]
                      (sents_labeled.uid, sents_labeled.label) == (code, label))]:
    print(sents_labeled.uid[i], sents_labeled.label[i])

for i in sents_labeled[(sents_labeled.label != sents_labeled.reference_label) &
                      (sents_labeled.uid, sents_labeled.label) != (sents_labeled.reference_uid, sents_labeled.reference_label)]:
    print(sents_labeled.uid[i], sents_labeled.label[i])
```

```
def evaluate_code(code, agg):
    print('Confusion Matrix')
    print(pd.crosstab([agg[code+'__l1m']>0, agg[code]>0]))
    print()
    print('Performance')
    for score in [precision_score, recall_score, f1_score]:
        print(score, name = score+agg[code]>0.agg[code+'__l1m']>0))

l1m=['l1']
agg=[agg.uid[agg[code+'__l1m']=0 & (agg[code]=0)].sort_values('uid').index:
code+'__l1m_keep']] , sentis_labeled,sentence[i])

l1m=['l1']
agg=[agg[agg[code+'__l1m']=0 & (agg[code]=0)].sort_values('uid').index:
code+'__l1m_keep']] , sentis_labeled,sentence[i])
```

4

Reference List

1. Rothwell J. *How Parenting and Self-Control Mediate the Link Between Social Media Use and Youth Mental Health*. Institute for Family Studies and Gallup; 2023.2.
2. Thulin EJ, Kusunoki Y, Kernsmith PD, et al. Longitudinal Effects of Electronic Dating Violence on Depressive Symptoms and Delinquent Behaviors Across Adolescence. *J Interpers Violence*. 2024;39(11-12):2526-2551. doi:10.1177/08862605231221281
3. John A, Glendenning AC, Marchant A, et al. Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review. *J Med Internet Res*. 2018;20(4):e129. doi:10.2196/jmir.9044
4. Macrynika N, Auad E, Menjivar J, Miranda R. Does social media use confer suicide risk? A systematic review of the evidence. *Comput Hum Behav Rep*. 2021;3:100094. doi:10.1016/j.chbr.2021.100094
5. Sumner SA, Ferguson B, Bason B, et al. Association of Online Risk Factors With Subsequent Youth Suicide-Related Behaviors in the US. *JAMA Netw Open*. 2021;4(9):e2125860. doi:10.1001/jamanetworkopen.2021.25860
6. Chu J, Ganson KT, Baker FC, et al. Screen time and suicidal behaviors among U.S. children 9–11 years old: A prospective cohort study. *Prev Med*. 2023;169:107452. doi:10.1016/j.ypmed.2023.107452
7. Masuda N, Kurahashi I, Onari H. Suicide Ideation of Individuals in Online Social Networks. Szolnoki A, ed. *PLoS ONE*. 2013;8(4):e62262. doi:10.1371/journal.pone.0062262
8. U.S. Department of Health and Human Services. *Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory*. U.S. Department of Health and Human Services; 2023:25. Accessed November 20, 2024. <https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/social-media/index.html>
9. Jaycox LH, Murphy ER, Zehr JL, Pearson JL, Avenevoli S. Social Media and Suicide Risk in Youth. *JAMA Netw Open*. 2024;7(10):e2441499. doi:10.1001/jamanetworkopen.2024.41499
10. Curtin S, Garnett M. *Suicide and Homicide Death Rates Among Youth and Young Adults Aged 10–24: United States, 2001–2021*. National Center for Health Statistics (U.S.); 2023. doi:10.15620/cdc:128423
11. Durkheim E. *Suicide*. 0 ed. Routledge; 2005. doi:10.4324/9780203994320
12. O'Connor RC, Kirtley OJ. The integrated motivational–volitional model of suicidal behaviour. *Philos Trans R Soc B Biol Sci*. 2018;373(1754):20170268. doi:10.1098/rstb.2017.0268
13. Lindley LC, Policastro CN, Dosch B, Ortiz Baco JG, Cao CQ. Artificial Intelligence and the National Violent Death Reporting System: A Rapid Review. *CIN Comput Inform Nurs*. 2024;42(5):369-376. doi:10.1097/CIN.0000000000001124
14. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can Large Language Models Transform Computational Social Science? Published online February 26, 2024.
15. Dubey A, Jauhri A, Pandey A, et al. The Llama 3 Herd of Models. Published online 2024. doi:10.48550/ARXIV.2407.21783