Area of Focus. 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. Recent rise in youth suicide[8] highlights the urgent need to understand how online experiences contribute to this public health issue.

Our mixed-methods approach responds to this challenge by proposing to develop four novel variables focused on instances when social norms or established behaviors were violated or significantly altered in online digital spaces. While NVDRS contains a 'Disclosed to Social Media' variable, it does not otherwise characterize how youth may engage in online spaces prior to suicide. The goals of this project are to 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.

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 Suicide[9] 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) theory[10] 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 (no suicidal ideation, but experiencing a trigger such as cyberbullying), motivational (experiencing suicidal ideation in response to trigger). 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. The entry into the volitional phase (suicidal action) may be recognized in behaviors that normalize death or increase access to injury means (e.g., discussing injury means online, exposure to violent content).

Methods: Identifying Key Themes. First, we performed a thematic analysis to define a set of four variables describing non-normative behaviors in online spaces. To obtain a sub-sample with relatively high density of relevant cases, a set of 19 keyphrases were developed to identify a subset of 1,142 narratives that are likely to reference the decedent's online behavior. These narratives contain words or phrases often used to describe online activity (e.g., "online" or "post") (Table 1). Eight authors conducted an initial exploratory review of 50 of these narratives, 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 capturing the nuances of the behaviors described in the narrative (e.g., "strong emotional reactions to content they consumed," "creating or consuming sexually explicit content"). A third step involved grouping these subcodes into four themes described below. These themes capture qualitatively and quantitatively distinct attributes about online behavior than what is already in the NVDRS variables; for instance, 93% of cases identified by the keyphrases were not already identified by the existing Disclosed to Social Media variable. Our next step will be to code an additional set of 300 narratives to refine these themes.

Types of Phrases	**# Cases**	**SM Disclosure = T**	**SM Disclosure = F**
2 Web Phrases *(online, cyber)*	99	3 (3%)	96 (97%)
8 Social Media Phrases *(facebook, etc.)*	330	33 (10%)	297 (90%)
2 Message Phrases *(post, chat)*	382	35 (9%)	347 (91%)
2 Format Phrases *(image, stream)*	28	4 (14%)	24 (86%)
5 Other Phrases *(gam, dating, porn, etc.)*	303	5 (2%)	298 (98%)
All Phrases	**1,142**	**80 (7%)**	**1,062 (93%)**

Table 1. Cases for annotation in thematic analysis, identified by 19 keyphrases for online spaces.

Proposed Variables. From our initial review, we propose constructing four novel variables, contextualized within the two theories (see Table 2). As shown in Table 2, these variables capture suicide risk factors that operate across diverse theoretical mechanisms. I. "Affective expression" refers to instances when decedents shared private or intimate content that exposed their vulnerability (e.g., identities, fears, negative emotional states, or self-harm; example: "V posted a photo on social media of his cut marks") or cases where the decedent was upset by content online ("V [...] got upset with his girlfriend about a photo she posted"). Sharing private emotional states within an online space or group that was not designed for, can be seen as transgression of norms of integration. These online behaviors map capture 'threat-to-self moderators' that facilitate a t*ransition from pre-motivation to motivational phase.*

	Definition	IMV	Integration / Regulation
Affective expression	shared private/intimate content, upset by content online	Pre-/Motivational	Hi / Lo
Cessation of online activity	voluntary/forced reduction in online activities	Motivational moderators	Hi / Lo
Exposure to or sharing of explicit content	engaged with content that is violent, offensive, or inflammatory	Volitional moderators	Lo / Lo
Victimization: violation of trust	online victimization resulting in breach of trust	Pre-motivational	Lo / Lo

Table 2. The four proposed novel variables in context of both theoretical frameworks.

II. "Withdrawal from or cessation of online activity," refers to abrupt or complete reduction of all online activities ("[V] had withdrawn the past month from friends and social media" or "V had been grounded from his phone"). Behaviors of withdrawal and disconnection indicate a *loss of integration*, while restrictions of access indicate *excess regulation*. These behaviors may be linked to *motivational moderators* that convey underlying themes of thwarted belongingness. III. "Exposure to or sharing of explicit content" captures descriptions of decedents engaged with or sharing content that is violent, offensive, or inflammatory ("[V was] seen watching videos online about people killing themselves"). These behaviors transgress *norms of regulation* and may be linked to *volitional moderators* that may desensitize someone to violent imagery and death. IV. "Victimization: violation of trust" encompasses narratives describing decedents' experiences of online victimization, which may involve cyberbullying, harassment, or unauthorized sharing of information ("[V had] previous issues with being bullied playing video games"). These behaviors suggest a lack of regulation and integration and serve as volitional moderators triggering increased suicidality. Our codes do not include cases where narratives indicate only that the decedent engaged in common online behaviors only: computer-mediated communication (e.g., texting or DMing a friend), use of social media, online gaming, or online dating platforms without further description of their online activities (e.g., someone used the decedent's social media accounts to track their location post-mortem), or online schooling.

Future Methods: LLM Classification of Narratives into Themes. Our next steps will be to explore the potential of LLMs to identify complex themes pertaining to youth suicide from NVDRS narratives. While prior literature finetuned LLMs to detect themes in the narratives,[11] 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.[12] We plan to test the zero-shot performance of LLMs in identifying our four variables.

Before the final submission, *we plan to craft prompts directing an open-sourced LLM like Llama3.1-8B-Instruct[13] to classify whether each narrative includes each theme*. As proof of concept, we prompted (proprietary) GPT-4o: "In the following narrative, does V share private or intimate content, fears, emotional states, or otherwise vulnerable information online?" This two-stage prompt goes on to instruct the LLM to create a structured output with 1) a binary indicator for the "affective expression" theme and 2) a list of relevant sentences. In preliminary explorations, giving an LLM this prompt produced high precision outputs (e.g., sentences like "V had recently come out as bi-sexual on social media" and "V posted a picture on Snapchat of himself with a rope around his neck"). To more systematically evaluate the zero-shot performance of LLMs at this task, we plan to manually annotate 500 narratives, 400 containing one or more of the keyphrases described previously and 100 that do not. We will test the precision and recall of the LLM-identified themes against these human annotations. We also hope to test the effects of different types of prompts (e.g., direct vs chain-of-thought, binary vs. multilabel classification) and different open-source LLMs (Llama3.1-8B-Instruct, GPT-j, Mistral-7B-Instruct-v0.3, etc.) on task performance. Generation of these variables will provide opportunities for future research into how experiences in social media may inform suicide in youth.

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. Macrynikola 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. 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
9. Durkheim E. *Suicide*. 0 ed. Routledge; 2005. doi:10.4324/9780203994320
10. 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
11. 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
12. 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.
13. Dubey A, Jauhri A, Pandey A, et al. The Llama 3 Herd of Models. Published online 2024. doi:10.48550/ARXIV.2407.21783