

Beyond the Rabbit Hole: Mapping the Relational Harms of QAnon Radicalization

Bich Ngoc Doan¹, Giuseppe Russo¹, Gianmarco De Francisci Morales²

¹EPFL, Lausanne, Switzerland ²CENTAI, Turin, Italy

bich.doan@epfl.ch, giuseppe.russo@epfl.ch, gdfm@acm.org

Abstract

Large-scale computational research on conspiracy theories has focused exclusively on believers’ online behavior, leaving the harm experienced by those closest to them under-examined. This paper bridges this gap by analyzing 12 747 stories from *r/QAnonCasualties*, an online support group for people who have “lost” someone to conspiracy beliefs. We design a computational pipeline to extract fine-grained thematic traits from personal narratives and cluster them into six coherent radicalization personas, which we then link to the emotional toll reported by narrators via LLM-assisted emotion detection and regression modeling. We find that personas are meaningful predictors of specific emotional harms: radicalization perceived as a deliberate ideological choice is associated with anger and disgust, while personas marked by personal and cognitive collapse correspond to fear and sadness. This work provides an empirically grounded computational framework for understanding the relational harms of radicalization, opening new avenues for research into its wider social consequences.

1 Introduction

Conspiracy theories pose a significant social and political challenge by driving polarization, eroding trust in public institutions (Van der Linden, 2015), and inciting real-world violence (Munn, 2021). As these theories proliferate online (Russo et al., 2023b), computational research has largely relied on social media analysis to map the diffusion of misinformation and track believers’ online footprint (Phadke et al., 2021; Wang et al., 2023a; Engel et al., 2022; Corso et al., 2025a).

While this paradigm yields valuable large-scale insights into the online evolution of conspiratorial communities, it has two key limitations. First, believer profiles are built primarily from online content, overlooking the offline characteristics that

help explain the psychological vulnerabilities and triggers underlying belief adoption. Second, focusing on believers alone underestimates the consequences that reach beyond the individual to affect their immediate social circle. Anecdotal studies report the harm endured by relatives and acquaintances of people falling into the “rabbit hole”: social erosion, depression, and the risk of adopting the same beliefs (Xiao et al., 2021; Engel et al., 2023; Mastroni and Mooney, 2024). Yet these consequences remain unmeasured at scale, leaving a major blind spot in the literature.

To address this gap, we turn to *r/QAnonCasualties*, a subreddit where users post first-hand stories about relatives or close acquaintances who have adopted QAnon and adjacent conspiracy beliefs—arguably the largest and most prevalent contemporary case study of mass radicalization—detailing the radicalization journey of the individual and the behavioral shifts observed throughout this process. These data offer a dual perspective: each post captures the offline radicalization trajectory of the relative or acquaintance—whom we refer to as the *object of narration* (OON)—as perceived by the post’s author, defined hereafter as the *narrator*, while simultaneously documenting the emotional and relational harm they experience. To study this dual perspective, we formulate two research questions:

RQ1: Which archetypes, or “personas,” emerge from the thematic traits perceived across the OON’s radicalization trajectories?

RQ2: How are these radicalization personas associated with the emotional toll experienced by the narrator?

To answer **RQ1**, we use an LLM co-annotation framework to extract thematic traits from narrative texts and organize them—drawing on radicalization

theory (Borum, 2011)—into three temporal phases spanning pre-existing conditions, trigger factors and post-radicalization characteristics. We then model the co-occurrence of these traits to derive six data-driven radicalization personas.

To address **RQ2**, we apply regression analysis to compositional data (via Isometric Log-Ratio transformation) to examine how exposure to these personas relates to narrators’ emotional outcomes. We link the OON’s radicalization personas to reported psychological distress and demonstrate that they are strong predictors of specific emotional harm.

Overall, we make three contributions:

- C1:** We release QAnonCasualties-12k, the first publicly available dataset built from 12 747 first-hand stories documenting offline radicalization into QAnon, annotated with thematic traits characterizing the OON’s radicalization trajectory and the narrator’s emotional response.¹
- C2:** We derive six data-driven *radicalization personas*, the first empirical typology of offline radicalization journeys as witnessed and reported by close relations.
- C3:** We provide large-scale quantitative evidence that radicalization personas predict specific emotional harms experienced by the narrator, thus reframing radicalization as a measurable relational phenomenon.

2 Related Work

Characterizing Online Radicalization. Computational research has extensively mapped online radicalization through digital traces. Several works study how users are socially drawn into extremist communities (Phadke et al., 2021), where echo chambers amplify divisive rhetoric (Samory and Mitra, 2018), violent content (Rousis et al., 2022), and emotionally-charged disinformation (Sharma et al., 2022; Marino et al., 2024). These communities also prove highly resilient, withstanding moderation by migrating to new platforms or spilling over into the mainstream (Russo et al., 2023b,a)

Radicalization, however, extends beyond online

spaces and reshapes the offline lives of its targets. Susceptibility to conspiracy thinking (Corso et al., 2025b) often builds on pre-existing vulnerabilities such as health struggles or institutional distrust (Xiao et al., 2021), and may culminate in mental health challenges (Engel et al., 2023). However, research documenting these dynamics relies mainly on small-scale interviews, which constrain generalizability. Our study addresses this by analyzing the narratives on r/QAnonCasualties, capturing the offline, observational dimensions of radicalization within a large corpus of first-person accounts.

Radicalization Framework. Radicalization is a complex, multi-faceted process that varies across populations (Wang et al., 2023a; Phadke et al., 2022). Several frameworks have been developed to model its components and progression. At a granular level, research identifies radicalization through discrete signals: computationally, via the adoption of specific jargon or hateful language (Fernandez et al., 2018; Rowe and Saif, 2016), and qualitatively, via offline behaviors such as pathological fixation on public figures or explicit threats (Reid Meloy et al., 2012). These signals are typically part of a broader, incremental process. Stage-based models further conceptualize radicalization as a temporal progression through phases, from grievance to indoctrination and action (Borum, 2011; Wolfowicz et al., 2021). We draw on these frameworks to operationalize radicalization as a trajectory of thematic traits, and examine how they coalesce into distinct personas (RQ1).

Radicalization Impacts Beyond the Individual. Radicalization harms not only believers but also their social circle. There is growing evidence that having a loved one involved in a conspiracy theory may lead to anxiety and PTSD (Moskalenko et al., 2023), strain relationships (Mastroni and Mooney, 2024), and require specialized psychosocial support (St-Amant et al., 2023). The nature of these harms likely depends on their underlying causes, yet little is known about the spread of emotional tolls across radicalization profiles. To fill this gap, our study links distinct radicalization personas to the form of distress with which they are associated (RQ2), thus providing the first large-scale test of this relation.

¹We release Reddit post identifiers alongside our derived annotations (thematic traits, persona mixtures, emotion labels, and narrator control variables) rather than raw post text, to protect user privacy and respect Reddit’s terms of service. Code and data are available at <https://github.com/ngoccc/qanoncasualties-analysis>.

3 QAnonCasualties-12k Dataset

3.1 Data Collection

We construct our dataset from 31 837 posts on r/QAnonCasualties, a subreddit where users document the radicalization of relatives or acquaintances into QAnon and adjacent conspiracies. We focus on accounts that detail the believer’s behavioral evolution over time. Following [Antoniak et al. \(2019\)](#), we preprocess the corpus by removing automated accounts, duplicates, and posts shorter than 50 words. We then apply a structural filter to retain only those that capture an interpersonal dynamic: ones that mention both a *narrator* (the post author) and an *object of narration* (OON, the believer). To operationalize this filter, we use a dependency parser ([Choi et al., 2015](#)) to extract nouns and pronouns, manually categorize the most frequent items into a narrator set (e.g., “I”, “me”) and an OON set (e.g., “mother”, “husband”), as listed in Table 1, and keep only posts containing at least one entity from each set.

Table 1: Entities and their representative n-grams.

Entity Category	n-grams
<i>Narrator</i>	
Narrator (self)	I, me, myself
<i>Object of Narration (OON)</i>	
Parent	mother, father, parent
Sibling	brother, sister, sibling, cousin
Extended family	aunt, uncle, grandparent
Partner	husband, wife, boyfriend, girlfriend
Friend	friend, roommate, colleague

Three annotators reviewed a random sample of 200 retained posts: 87% satisfy the dual-perspective requirement (Fleiss’ $\kappa = 0.71$). The pipeline yields a final dataset of 12 747 posts (additional details in Appendix A.1).

3.2 Data Processing

Posts on r/QAnonCasualties are unstructured: they mix biographical detail, affective testimony, and second-hand reporting, complicating the extraction of OON-specific attributes needed to reconstruct radicalization trajectories. To address this, we apply BERTopic ([Grootendorst, 2022](#)) to extract topical patterns from post sentences, using UMAP for dimensionality reduction and HDBSCAN for clustering (hyperparameters in Appendix A.3). This process yields 461 topics. We benchmark BERTopic against LDA ([Blei et al., 2003](#)) and prompt-based LLM topic modeling ([Wang et al.,](#)

[2023b](#)) (Appendix A.2), and find that BERTopic achieves higher topic coherence, consistent with prior findings ([Egger and Yu, 2022](#); [Wang et al., 2023b](#)).

Manual inspection revealed that many of the 461 topics were not descriptive of OON attributes (e.g., trivial life details, conversational artifacts). To filter them, we develop an **LLM Co-annotation Framework**: a human-in-the-loop pipeline combining LLM-assisted topic filtering with expert validation. We curate few-shot exemplars of relevant and irrelevant topics—each with topic keywords, representative posts, a binary label, and a short rationale—and prompt gpt-4o-mini to classify each remaining topic, marking it relevant if it pertains to (i) the OON’s belief adoption, (ii) behavioral change, or (iii) resulting interpersonal dynamics. The first author independently annotated all topics; disagreements with the LLM were resolved in consultation with the two co-authors, yielding a final set of 198 relevant topics, each subsequently labeled with a concise LLM-generated descriptor.

4 RQ1: Trajectories and Personas

4.1 Radicalization Trajectories

To characterize the OON’s radicalization trajectories, we map the 198 OON-relevant topics onto the staged model of radicalization proposed by [Borum \(2011\)](#) and operationalized for empirical trajectory analysis by [Klausen \(2016\)](#). The model conceptualizes radicalization as a progression organized around a central catalyzing event (the *trigger*) that mediates between an individual’s antecedent vulnerabilities and their subsequent ideological and behavioral transformation. This framing aligns naturally with our setting: testimonies on r/QAnonCasualties are typically structured along this temporal arc, recounting who their loved one was before (*Pre-Radicalization Conditions*), what happened (*Trigger*), and what they became after (*Post-Radicalization Characteristics*)—the three phases we adopt as the backbone of our analysis.

Each LLM-co-annotated topic is assigned to one of these phases by the first author. Assignments were validated by the co-authors, with disagreements resolved through discussion. We then manually merge conceptually similar topics into higher-level themes (Appendix A.4), yielding a final set of 50 thematic *traits* of radicalization trajectories. Each post is thus represented as a *profile*: the subset of

traits it expresses.

Figure 1 visualizes the resulting trajectory space as a flow across the three phases, with edge widths proportional to the frequency of co-occurring traits within profiles. Three findings stand out:

Radicalization antecedents are ideologically anchored but individually heterogeneous.

Pre-Radicalization Conditions span psychological (Physical Health Disorders, Mental Health Disorders), socio-economic (Financial Struggles), and demographic dimensions. Yet, Conservative Political Identity stands out as the dominant antecedent, appearing in 1489 profiles—nearly 50% more than the next most common factor. This suggests that, despite substantial heterogeneity in vulnerabilities, conservative ideological orientation is the strongest precursor to QAnon adoption in our corpus.

Triggers concentrate around a pandemic shock and its media ecosystem.

Trigger Factors are strikingly concentrated: COVID-19 Lockdowns (2402) dominate by a wide margin, with the remainder clustering around Religious Influence (1181) and platform-specific media exposure (Fox News, Facebook, YouTube Algorithms). Political events, by contrast, play a comparatively minor role (Trump's Election, 181 profiles), pointing instead to the pandemic information environment as a potentially central driver of recruitment. Nevertheless, a considerable share of profiles (1714) lack a clearly identifiable trigger (Unknown), suggesting the catalyzing moment is often internal or simply hidden from view.

Post-Radicalization Characteristics cluster along two largely independent axes.

Outcomes separate cleanly into a *relational* axis (Social Deterioration and Interpersonal Harm) characterized by communication tension, isolation, and estrangement, together with shifts in Lifestyle and Media Use, and an *ideological* axis (adoption of core QAnon tenets centered on Child Harm and Apocalyptic Narratives), interwoven with Pro-Trump and War-Inducing Rhetoric, and accompanied by a broader rejection of established knowledge via Medical and Science and Technology Mistrust. The two axes co-occur in only 32.6% of profiles, thus suggesting that *behavioral* and *epistemic* radicalization are partially decoupled: a distinction

that mirrors the two-pyramids account separating radicalization of action from radicalization of opinion (McCauley and Moskalenko, 2017).

4.2 Clustering Radicalization Profiles

Building on the trajectory framework, we now ask whether the OON profiles cluster into a small number of recurring *radicalization personas*, which we define as latent clusters capturing co-occurring sets of traits across profiles.

A key methodological assumption is that real-world radicalization profiles are unlikely to belong to a single archetype: a believer may, for instance, combine pandemic-era distrust of institutions with a pre-existing right-wing political identity and a religious worldview. Hard-clustering methods such as *k*-means, which force each profile into a single cluster, would obscure precisely the kind of overlap we expect to find. We therefore adopt a probabilistic, mixed-membership formulation, in which each profile is represented as a distribution over personas and each persona as a distribution over traits. Following Yang et al. (2016), who use a similar formulation to derive behavioral archetypes, we instantiate this with Latent Dirichlet Allocation (LDA), representing each profile as a bag-of-traits.

To select the number of personas, we sweep *k* from 4 to 15 and select *k* via topic coherence (Appendix A.6 and A.7). The metric peaks at *k* = 6, and a subsequent qualitative review of the resulting personas confirms that this value yields the most distinct and coherent archetypes. We use *k* = 6 henceforth.

4.3 Persona Analysis

Our six personas (**Health-Triggered Conspiracy Theorist**, **Political Extremist**, **Social Media Spiral**, **Religious Apocalypticist**, **Conservative Identity Protector**, and **Pandemic-Triggered Skeptic**) are described in Table 2. Per-persona narratives and statistics are in Appendices A.8 and A.9. Overall, these personas are not artifacts of hard partitioning: profiles occupy on average 1.70 personas (threshold = $1/k$); 52.2% exhibit mixed membership, confirming that the probabilistic formulation captures genuine overlap rather than serving as a methodological formality.

We frame the contribution of this typology in two parts. First, **we provide the first integrated empirical typology of QAnon believers from wit-**

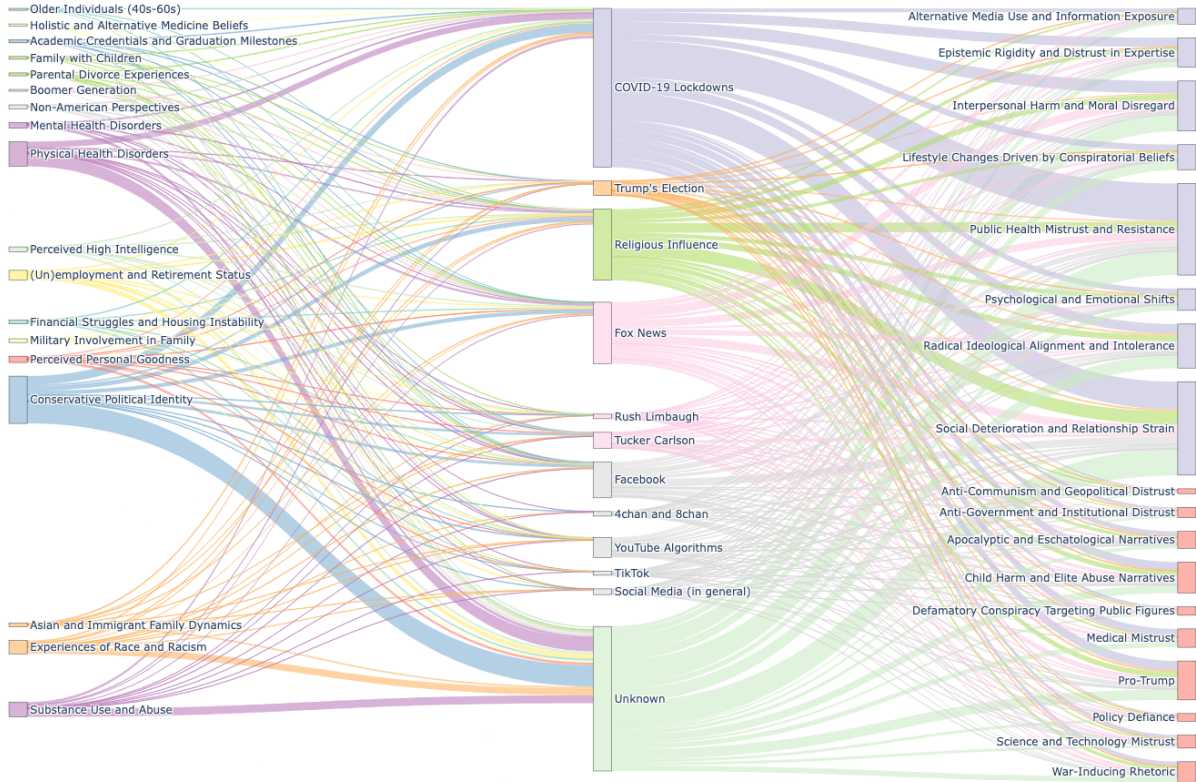


Figure 1: Flow of 50 thematic traits across the three radicalization phases.

nessed testimony. Prior work has identified individual mechanisms in isolation: right-wing media ecosystems (Sipka et al., 2022), algorithmic amplification (Hannah, 2021), Christian eschatology (Miotto and Droogan, 2025), COVID-19 conspiracy convergence (Morelock and Narita, 2022), and personal health attribution to conspiratorial actors (Xiao et al., 2021). However, it has not shown how these mechanisms intertwine within a believer population. Each of our six personas instantiates one of these previously documented mechanisms, and their relative frequencies (Table 2) provide the first quantitative estimate of their respective prevalence in the QAnon-adjacent population observable through r/QAnonCasualties. Second, **mixed membership reveals interaction effects invisible to single-mechanism studies.** The well-documented fusion of Christian symbolism with Trumpism (Miotto and Droogan, 2025) surfaces as co-loading on the Religious Apocalypticist and Political Extremist personas in over 1000 believers in our corpus. Mixed membership thus reveals not just who radicalized, but how distinct pathways converge and reinforce one another.

5 RQ2: Persona-Specific Emotional Harm

Our second goal is to quantify the impact that different radicalization personas have on the emotional toll experienced by their relatives and acquaintances, i.e., to model the relationship between the *object of narration* (OON) and the *narrator*.

5.1 Regression Setup

We model this relationship via a binary logistic regression. For each of four negative emotions $y \in \{\text{anger, fear, sadness, disgust}\}$, we fit

$$\text{logit } \Pr(y = 1 \mid \mathbf{x}, \mathbf{c}) = \beta_0 + \beta^\top \mathbf{x} + \gamma^\top \mathbf{c}, \quad (1)$$

where y is the presence of the emotion in the narrative, $\mathbf{x}$ encodes the OON’s persona composition, and $\mathbf{c}$ is a vector of narrator-level controls. We describe each component in turn below.

Dependent variables: narrator emotions (y). We operationalize the emotional toll by using four discrete negative emotions from Plutchik’s framework: anger, fear, sadness, and disgust (Plutchik, 2001). We choose this categorical approach over dimensional models (e.g., valence and arousal) for its interpretability, as identifying specific emotions

Table 2: The six radicalization personas, their core descriptions, and their dominant trajectory sub-diagrams.

No.	Persona	Description	Sub-Sankey Diagram
P1	The Health-Triggered Conspiracy Theorist	This persona’s radicalization begins with a personal health crisis or substance abuse issue. Their initial mistrust of doctors evolves into a comprehensive conspiratorial worldview that eventually incorporates QAnon. This pathway is marked by severe social and professional collapse, including psychotic episodes, aggressive behavior, and the destruction of their former relationships and career.	
P2	The Political Extremist	Driven by a pro-Trump identity and a diet of right-wing media, this persona adopts a fierce “us versus them” mentality. They view political opponents as an existential threat, justifying extreme rhetoric and the destruction of personal relationships over political disagreements. While they may carefully vet sources within their media bubble, they remain trapped in an echo chamber that reinforces their apocalyptic fears.	
P3	The Social Media Spiral	This persona falls down algorithmic rabbit holes on platforms such as YouTube and 4chan. What starts as a simple search for alternative information becomes a spiral of increasingly extreme content, culminating in an obsession with child harm narratives. They lose the ability to think critically and become highly susceptible to the influence of online influencers, while rejecting mainstream sources.	
P4	The Religious Apocalypticist	This persona fuses conservative Christianity with political conspiracy, interpreting current events through the lens of Biblical end-times prophecy. QAnon is seen as proof of a divine spiritual war, creating a powerful, self-reinforcing belief system. This worldview often includes a fear of technology as the “mark of the beast” and a conviction that a final battle between good and evil is imminent.	
P5	The Conservative Identity Protector	This persona becomes radicalized in what they see as a defense of traditional conservative values against cultural threats, such as changing norms around race, gender, and sexuality (e.g., LGBTQ+ movements). Their identity becomes so paramount that they view disagreement as a personal betrayal, leading to an intolerance for opposing views and a willingness to prioritize ideological purity over family relationships.	
P6	The Pandemic-Triggered Skeptic	Ignited by the COVID-19 pandemic, this persona’s pre-existing skepticism of mainstream medicine and preference for holistic health made them vulnerable to anti-vaccine messaging. This initial doubt quickly expands into broader government control conspiracies and a complete rejection of medical and scientific authority.	

such as “anger” versus “sadness” provides more actionable insights into the nature of interpersonal harm than abstract affective scores.

Extracting these emotions from our corpus is non-trivial: posts are long-form, multi-actor, and describe highly specific lived experiences. Standard approaches, such as lexicon-based methods and pre-trained classifiers, fail to disentangle the narrator’s emotions from those of the people they de-

scribe. Training a custom classifier would require costly manual annotation beyond the scope of this work. We therefore adopt an LLM-assisted strategy using gpt-4o-mini, selected for its strong instruction-following capabilities and cost efficiency. Prompts are iteratively designed and evaluated following prior LLM-based emotion detection work (Greschner and Klinger, 2024; Liu et al., 2024; Niu et al., 2024). Empirical validation on a

100-post sample confirms this choice: GPT-based annotation achieves mean $\kappa = 0.6$ against two independent annotators, substantially outperforming both NRC (Mohammad and Turney, 2013) and GoEmotions-finetuned (Lowe, 2022) baselines (full results and error analysis in Appendix A.10). The final prompt (Appendix A.5.4) is applied to the full dataset, yielding four binary dependent variables.

Controls: narrator attributes (c). To isolate the effect of persona composition, we control for three narrator-level attributes: the narrator’s relationship to the OON, and their inferred gender and age. We extract these attributes via a hybrid annotation pipeline standard in demographic inference (Jagfeld et al., 2021; Rosario, 2023; Harvey et al., 2025). First, we apply high-precision regular expressions adapted from Tiginova et al., 2020 and our entity n-grams Table 1 (full details in Appendix A.11) to identify explicit self-disclosures (e.g., “my husband”, “I [35F]”). We then benchmark zero-shot classification with gpt-4o-mini against this set, after masking the explicit disclosures used as ground truth. The model achieves F1 score of 0.95 for relationships, 0.77 for age, and 0.70 for gender. Based on this validation, we use the LLM to infer the attributes for posts where they are not explicitly stated.

Predictors: persona composition (x). The predictors of interest are the six radicalization personas identified in Section 4.3. For each post (i.e., profile), the LDA model estimates a probability distribution over the six personas, which would naturally serve as x . However, plugging these probabilities directly into Equation (1) is statistically invalid, and a transformation is required.

Handling Compositional Predictors via ILR. The regression cannot be fit on the raw persona probabilities: they are non-negative and sum to one (*compositional data* (Aitchison, 1982)), which violates the assumptions of standard linear regression. We address this issue via an *Isometric Log-Ratio* (ILR) transformation (Egozcue et al., 2003), which maps the six probabilities into five orthogonal coordinates called *balances*, each capturing the relative dominance of one group of personas (the *numerator*) over another (the *denominator*). The grouping is defined by a Sequential Binary Partition (SBP) that reflects thematic distinctions across the personas: the top-level balance contrasts

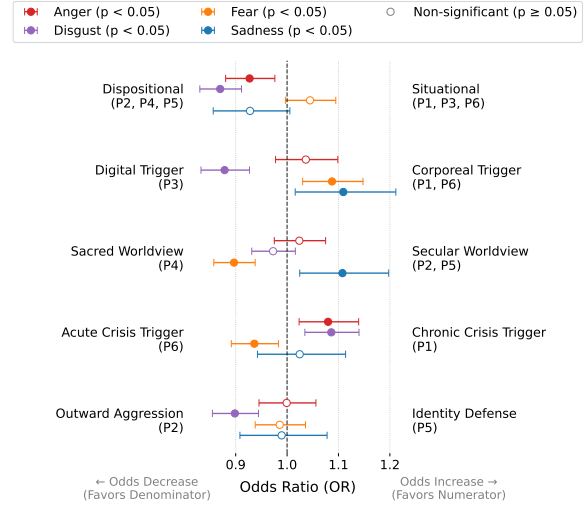


Figure 2: Effects of personas on narrator emotions. Points show odds ratios (OR) from logistic regression models, with bars indicating 95% confidence intervals.

Situational personas (radicalization stemming from external events such as a health crisis or algorithm) against *Dispositional* ones (rooted in pre-existing identity or faith), with four further balances encoding nested contrasts within each group. The full SBP and ILR formulation is shown in Table 3.

5.2 Regression Results

We fit one logistic regression per emotion (Equation (1)) and report results as odds ratios (OR): $OR > 1$ indicates higher odds of the emotion with greater dominance of the numerator group, whereas $OR < 1$ indicates higher odds with the denominator group. We interpret findings at two levels: (1) the high-level conceptual themes of each balance, and (2) the defining persona characteristics within those balances. The results, visualized in Figure 2, reveal two distinct emotional patterns.

Anger and disgust are most strongly associated with radicalization perceived as an active process inflicting tangible harm rather than a passive misfortune, mirroring psychological research linking anger to perceived intentionality (Weiner, 1985; Vasilopoulos et al., 2019). This pattern emerges in two key balances. First, our primary balance shows that profiles dominated by *Dispositional* personas (rooted in pre-existing political or religious beliefs) are more likely to evoke anger ($OR = 0.93$) and disgust ($OR = 0.87$) than those driven by *Situational* factors. Balance 4 refines this pattern within the Situational group: the *Chronic Personal Crisis* persona (P1, characterized by destructive behaviors) is linked to more

Table 3: Sequential Binary Partition (SBP) and Interpretation of the ILR Balances.

Balance #	Conceptual Grouping and Rationale	Numerator Group	Denominator Group
1	<i>Situational vs. Dispositional</i> : This balance separates personas where radicalization is attributed to external, situational factors (e.g., a health crisis, algorithm) from those attributed to one's internal, pre-existing disposition (e.g., their political identity or religious faith).	P1 (Health), P3 (Social Media), P6 (Pandemic)	P2 (Political), P4 (Religious), P5 (Conservative)
2	<i>Corporeal vs. Digital Trigger</i> : Within the Situational group, this balance distinguishes between entry points rooted in the physical body and health (Corporeal) versus those rooted in the abstract, algorithmic world of online platforms (Digital).	P1 (Health), P6 (Pandemic)	P3 (Social Media)
3	<i>Secular vs. Sacred Worldview</i> : Within the Dispositional group, this balance separates radicalization based on modern political identities (Secular) from one based on eschatological, faith-based narratives (Sacred).	P2 (Political), P5 (Conservative)	P4 (Religious)
4	<i>Chronic vs. Acute Crisis</i> : Within the Corporeal group, this balance separates radicalization stemming from a long-term, chronic personal crisis (P1) from one triggered by a specific, acute public crisis (P6).	P1 (Health)	P6 (Pandemic)
5	<i>Inward Identity Defense vs. Outward Aggression</i> : Within the Secular group, this balance separates an inward-looking identity defense (protecting conservative values) from explicit outward aggression ("us vs. them" political combat).	P5 (Conservative)	P2 (Political)

anger (OR= 1.08) and disgust (OR= 1.09) than the *Acute Public Crisis* (P6, defined by rejection of scientific authority). Even when the initial trigger is external, directly harmful personal actions are seen as a more potent transgression than a more indirect, anti-authority stance.

Disgust, in particular, is sensitive to the specific moral content of the transgression (Chapman and Anderson, 2013). Balance 5 reveals that the *Outward Aggression* persona (P2, defined by political extremism) corresponds to stronger disgust (OR = 0.90) than the *Inward Identity Defense* persona (P5, focused on identity-based anxieties). Likewise, Balance 2 shows that the *Digital-based Social Media Spiral* (P3), defined by an obsession with abhorrent child-harm narratives, exhibits higher odds of disgust (OR = 0.88) than the *Corporeal-based* personas (P1 and P6), which reflect more generalized personal and health-related deterioration. These findings highlight that disgust may function as a targeted moral response to perceived ideological depravity, rather than as a generic reaction to harm.

Fear and sadness are linked to grieving a perceived cognitive and personal collapse—a form of “ambiguous loss” (Pauline and Boss, 2009) where narrators mourn a loved one’s profound psychological transformation and loss of identity. The tangible evidence of a loved one’s physical and

mental decline in the *Corporeal-based* personas (P1, P6) has a stronger association with both fear (OR = 1.09) and sadness (OR = 1.11) than the more abstract *Digital-based* projection of P3 (Balance 2). Moreover, the ideological framing of the radicalization reveals a divergent impact on these emotions. Balance 3 shows fear to be more prominent in narratives of the *Sacred Religious Apocalypticist* (P4) (OR = 0.90), while sadness is more prevalent in those of *Secular* personas (P2, P5) (OR = 1.11). This nuanced result suggests that while religious detachment from reality is profoundly frightening, narrators perceive the hostility-induced transformation of secular extremism as a more saddening loss of identity—echoing the interpersonal grief documented in studies of affective polarization (Iyengar et al., 2019).

6 Discussion and Conclusion

This study introduced and validated a computational framework for moving beyond the individual conspiracy believer to model the offline, relational consequences of online radicalization. By analyzing thousands of personal stories, we first developed a pipeline to map radicalization trajectories and derive six recurring radicalization personas. We then provided the first large-scale, quantitative evidence that these computationally-derived personas are strong predictors of the specific emo-

tional harms inflicted on loved ones.

A core strength of our multi-stage pipeline—combining contextual topic modeling, graphical modeling, and LLM-assisted classification—is its ability to extract multiple components of narrative structure from unstructured text (Piper et al., 2021). Its modular design makes it adaptable to other narrative-driven phenomena—such as other belief-based communities or support group dynamics, where similar pipeline stages could be selectively applied to extract narrative components of interest. Such adaptation would require re-deriving the domain-specific components such as entity n-grams, thematic traits, and LLM prompts for the new setting, while the pipeline’s architecture, from temporal mapping through persona aggregation to outcome regression, remains directly reusable.

Beyond methodology, the six personas represent a foundational feature set: by capturing stable, archetypal patterns in how radicalization unfolds, they open avenues for predictive modeling. For example, future work could explore the use of these personas to forecast critical outcomes such as relationship severance, the likelihood of offline violence, or one’s receptiveness to de-radicalization. In doing so, it takes a step toward measuring not just who radicalizes, but also what radicalization leaves behind.

Limitations and Future Work

Our study has several limitations that highlight avenues for future work. First, our operationalization of emotional toll, while providing a clear and interpretable starting point, relies on pre-defined negative emotions. This simplification inevitably overlooks more complex, context-specific feelings prevalent, such as helplessness, irony, or betrayal. This shortcoming lays the groundwork for future research to develop a more nuanced data-driven taxonomy of the emotional harms associated with radicalization. In addition, our reliance on LLMs paired with limited human validation represents a trade-off between scale and precision, introducing errors from both sides: from the LLM itself, and from the narrow scope of human review, which was largely disagreement-triggered rather than exhaustive. A promising avenue is to scale up human validation, which would also provide a stronger empirical basis for developing classifiers tailored to radicalization-related narratives,

better able to capture subtle linguistic markers of trauma and support-seeking than general-purpose models. Finally, the generalizability of our findings is bounded by our dataset’s specificity. Drawing exclusively from r/QAnonCasualties introduces self-selection bias, where narrators seeking online peer support may not represent all those affected by radicalization, influencing which pathways appear more or less common in our analysis. Future work should address this along three directions: extending to other platforms and media to mitigate selection effects, analyzing data longitudinally to model how radicalization trajectories evolve over time, and testing whether our personas generalize across other conspiracy movements beyond QAnon.

Ethical Consideration

The data used in this study were sourced from the public subreddit r/QAnonCasualties. Our collection and analysis of this data for non-commercial research purposes are consistent with Reddit’s Terms of Service. While the content is publicly accessible, we take additional steps to protect the privacy of both the narrators (the post authors) and the objects of narration (the radicalized individuals they describe). All data was anonymized, and we removed any personally identifiable information (PII) such as usernames. All quantitative results presented in this paper are in an aggregated form. We report on collective patterns (e.g., persona distributions, odds ratios). The anonymized text fragments sent to the LLM APIs for annotation contained no personal information. We complied with the terms of service of the LLM provider for all analyses.

References

- John Aitchison. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.
- Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Randy Borum. 2011. Radicalization into violent extremism ii: A review of conceptual models and empirical research. *Journal of strategic security*, 4(4):37–62.
- Nina Cesare, Christan Grant, Quynh Nguyen, Hedwig Lee, and Elaine O Nsoesie. 2017. How well can ma-

- chine learning predict demographics of social media users? *arXiv preprint arXiv:1702.01807*.
- Hanah A Chapman and Adam K Anderson. 2013. Things rank and gross in nature: a review and synthesis of moral disgust. *Psychological bulletin*, 139(2):300.
- Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396.
- Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2025a. [Conspiracy theories and where to find them on TikTok](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 8346–8362, Vienna, Austria. Association for Computational Linguistics.
- Francesco Corso, Giuseppe Russo, Francesco Pierri, and Gianmarco De Francisci Morales. 2025b. Early linguistic fingerprints of online users who engage with conspiracy communities. *arXiv preprint arXiv:2506.05086*.
- Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498.
- Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3):279–300.
- Kristen Engel, Yiqing Hua, Taixiang Zeng, and Mor Naaman. 2022. Characterizing reddit participation of users who engage in the qanon conspiracy theories. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22.
- Kristen Engel, Shruti Phadke, and Tanushree Mitra. 2023. Learning from the ex-believers: individuals’ journeys in and out of conspiracy theories online. *Proceedings of the ACM on human-computer interaction*, 7(CSCW2):1–37.
- Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.
- Miriam Fernandez, Moizzah Asif, and Harith Alani. 2018. Understanding the roots of radicalisation on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 1–10.
- Lynn Greschner and Roman Klinger. 2024. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and llms. *arXiv preprint arXiv:2412.15993*.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Matthew N Hannah. 2021. A conspiracy of data: Qanon, social media, and information visualization. *Social Media+ Society*, 7(3):20563051211036064.
- Daisy Harvey, Paul Rayson, Fiona Lobban, Jasper Palmier-Claus, Clare Dolman, Anne Chataigné, Steven Jones, and 1 others. 2025. Using natural language processing methods to build the hypersexuality in bipolar reddit corpus: infodemiology study of reddit. *JMIR infodemiology*, 5(1):e65632.
- Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22(1):129–146.
- Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven H Jones. 2021. Understanding who uses reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. *arXiv preprint arXiv:2104.11612*.
- Jytte Klausen. 2016. *A behavioral study of the radicalization trajectories of american "homegrown" al Qaeda-inspired terrorist offenders*. Brandeis University.
- Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.
- Sam Lowe. 2022. [Roberta-base model on goemotions](#).
- Erik Bran Marino, Jesus M Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(11):603.
- Lauren Mastroni and Robyn Mooney. 2024. “i one-hundred thousand percent blame it on qanon”: The impact of qanon belief on interpersonal relationships. *Journal of Social and Personal Relationships*, 41(9):2478–2499.
- Clark McCauley and Sophia Moskalenko. 2017. Understanding political radicalization: The two-pyramids model. *American Psychologist*, 72(3):205.
- Nicolò Miotto and Julian Droogan. 2025. ‘stand against the wiles of the devil’: Interpreting qanon as a pseudo-christian extremist movement. *Critical Sociology*, 51(3):503–526.
- Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.
- Jeremiah Morelock and Felipe Ziotti Narita. 2022. The nexus of qanon and covid-19: Legitimation crisis and epistemic crisis. *Critical Sociology*, 48(6):1005–1024.
- Sophia Moskalenko, BS Burton, J Fernández-Garayzábal González, and MM Bloom. 2023. Second-hand conspiracy theories: The social, emotional and

- political tolls on loved ones of qanon followers. *Democracy and Security*, 19(3):231–250.
- Luke Munn. 2021. More than a mob: Parler as preparatory media for the us capitol storming. *First Monday*.
- Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *arXiv preprint arXiv:2408.17026*.
- Boss Pauline and Pauline Boss. 2009. *Ambiguous loss: Learning to live with unresolved grief*. Harvard University Press.
- Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. What makes people join conspiracy communities? role of social factors in conspiracy engagement. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3).
- Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2022. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 770–781.
- Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.
- J Reid Meloy, Jens Hoffmann, Angela Guldemann, and David James. 2012. The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral sciences & the law*, 30(3):256–279.
- Charlotte Rosario. 2023. Age-specific linguistic features of depression via social media. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 33–43.
- Gregory J Rousis, F Dan Richard, and Dong-Yuan Debbie Wang. 2022. The truth is out there: The prevalence of conspiracy theory use by radical violent extremist organizations. *Terrorism and Political Violence*, 34(8):1739–1757.
- Matthew Rowe and Hassan Saif. 2016. Mining pro-is radicalisation signals from social media users. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 329–338.
- Giuseppe Russo, Manoel Horta Ribeiro, Giona Casiraghi, and Luca Verginer. 2023a. Understanding online migration decisions following the banning of radical communities. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 251–259.
- Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2023b. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 742–753.
- Mattia Samory and Tanushree Mitra. 2018. ‘the government spies using our webcams’ the language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24.
- Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Characterizing online engagement with disinformation and conspiracies in the 2020 us presidential election. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 908–919.
- Andrea Sipka, Aniko Hannak, and Aleksandra Urman. 2022. Comparing the language of qanon-related content on parler, gab, and twitter. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 411–421.
- Michèle St-Amant, Michael King, Laura Stolte, and Marc-André Argentino. 2023. Qollateral: The impact of qanon on loved ones and the potential for p/cve programs to help. *Journal for Deradicalization*, (35):33–75.
- Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. Reddust: a large reusable dataset of reddit user traits. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6118–6126.
- Sander Van der Linden. 2015. The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and individual differences*, 87:171–173.
- Pavlos Vasilopoulos, George E Marcus, Nicholas A Valentino, and Martial Foucault. 2019. Fear, anger, and voting for the far right: Evidence from the november 13, 2015 paris terror attacks. *Political Psychology*, 40(4):679–704.
- Emily L Wang, Luca Luceri, Francesco Pierri, and Emilio Ferrara. 2023a. Identifying and characterizing behavioral classes of radicalization within the qanon conspiracy on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 890–901.
- Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023b. Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241. IEEE.
- Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4):548.
- Rebecca Willis. 2019. The use of composite narratives to present interview findings. *Qualitative research*, 19(4):471–480.

Michael Wolfowicz, Yael Litmanovitz, David Weisburd, and Badi Hasisi. 2021. Cognitive and behavioral radicalization: A systematic review of the putative risk and protective factors. *Campbell Systematic Reviews*, 17(3):e1174.

Sijia Xiao, Coye Cheshire, and Amy Bruckman. 2021. Sensemaking and the chemtrail conspiracy on the internet: Insights from believers and ex-believers. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–28.

Diya Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in wikipedia. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 446–455.

A Appendix

A.1 Data Preparation Details

Here we elaborate on the data preparation steps summarized in the main text. The data obtained covers the period between its launch on July 3, 2019, and December 31, 2024. In the primary filtering stage, we exclude moderator and bot content by removing posts with the distinguished field set to *moderator*, authored by the AutoModerator, or containing polls (via the *poll_data* field). We also exclude posts self-labeled with non-narrative flairs such as ‘Off-topic’ or ‘Media Request’. Duplicate posts are identified and removed by exact match on post text (*selftext* field). We choose the 50-word length threshold for removing short posts after examining the overall post length distribution (see Figure A.1), where we manually audit samples at different lengths to ensure the cutoff removes trivial content while retaining concise but complete narratives.

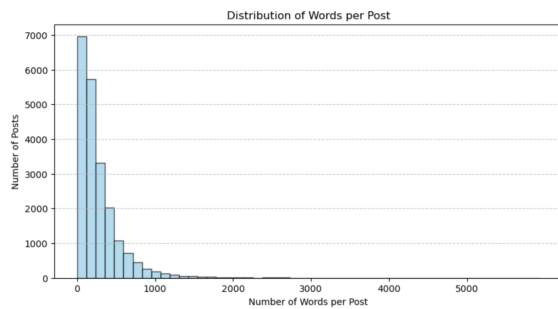


Figure A.1: Distribution of post lengths (word counts).

The second stage extracts entities to enforce the dual-perspective structure described in the main text. While Table 1 reports representative n-grams for conciseness, Table A.1 provides the complete set of n-grams used for filtering, including spelling

variants and synonyms (e.g., “mom”, “mum”, and “mother” are all mapped to the Parent category).

Table A.1: Entities and their full n-grams.

Entity	n-grams
Narrator (self)	I, me, myself
Parent	mom, mum, mother, dad, father, parent
Sibling	brother, sister, sibling, cousin
Extended family	grandma, grandpa, grandfather, grandmother, grandparent, aunt, uncle
Partner	husband, wife, boyfriend, girlfriend, spouse, fiancé, partner
Friends	friend, roommate, colleague

A.2 BERTopic Model Selection

We conduct an empirical comparison of three topic modeling approaches for our dataset: BERTopic, LDA, and an LLM-based method following the PromptTopic pipeline (Wang et al., 2023b) with gpt-4o-mini. All models are evaluated on the full filtered set of 12 747 posts with $k = 50$ topics fixed across models. We assess topic quality via NPMI coherence and diversity, and stability via Adjusted Rand Index (ARI) across three independent runs.

As shown in Table A.2, BERTopic outperforms both alternatives across all metrics. Most notably, its stability (ARI = 0.87) substantially exceeds that of LDA (0.14) and the LLM-based approach (0.53), indicating that BERTopic produces considerably more consistent topic assignments—a critical property for downstream persona derivation. We therefore proceed with BERTopic and tune its hyperparameters as described below.

Table A.2: Topic quality and stability across modeling approaches ($k = 50$, three independent runs).

Method	Diversity	Coherence	ARI
BERTopic	0.700	-0.030	0.873
LDA	0.317	-0.064	0.135
LLM-based	0.092	-0.136	0.533

A.3 BERTopic Configuration

We begin with BERTopic’s default UMAP and HDBSCAN parameters and tune iteratively, evaluating topic quality via hierarchical clustering dendrograms and intertopic distance maps. We observe that default settings produce coherent but overlapping topics—reflected in densely clustered regions of the intertopic distance map—and incrementally increase *min_cluster_size* until overlap is substantially reduced. The research team assessed coherence at each iteration, with disagreements resolved by consensus. The final configuration sets HDBSCAN’s *min_cluster_size* to 50, with all

other parameters at their defaults (see Table A.3). HDBSCAN identified 128 667 outlier sentences (35.3%), which were discarded, retaining 235 694 sentences for subsequent analysis.

Table A.3: BERTopic configuration parameters.

Parameter	Value
UMAP n_neighbors	15
UMAP n_components	5
UMAP metric	cosine
HDBSCAN min_cluster_size	50
HDBSCAN metric	euclidean

A.4 Merging Radicalization Topics

In the first research question, where we examine radicalization topics, we perform an aggregation process that consolidates 198 initial BERTopic outputs into 50 final themes used for the trajectory analysis. The rationale behind this choice is that when assigning the 198 topics to our three-phase framework, we observe a strong concentration in Phase 3 (*Post-Radicalization Characteristics*). This distribution likely reflects a narrative focus on the tangible, observable consequences of radicalization, which are often more salient to the storyteller than the specific pre-existing conditions or triggers. While this data distribution is itself an interesting finding, the high density of fine-grained topics in Phase 3 poses a challenge for creating a clear and interpretable visualization of the radicalization trajectories. A Sankey diagram with a disproportionately large and granular final stage would obscure the broader patterns of flow.

To address this and improve analytical clarity, we perform the thematic aggregation step, focusing on the Phase 3 topics. The full hierarchical structure of this consolidation for Phase 3 is presented in Table A.4. We additionally introduced "Unknown" as a label during visualization to represent profiles whose extracted topics contain no Phase 2 (Trigger) content, i.e. only Phase 1 and Phase 3 topics were identified. The main paper's Sankey diagram (Figure 1) displays this along with the final consolidated themes.

A.5 LLM Prompts

This section details all the prompts we used with the gpt-4o-mini for LLM annotation.

A.5.1 Radicalization Topic Relevance Classification

You are a qualitative coding assistant helping analyze radicalization pathways from r/QAnonCasualties posts.

Determine whether the following topic is relevant to understanding the Object of Narration's (OON's) radicalization pathway.

Give Label: Yes/No and Reason.

{few_shots}

Topic representation: {representation}
Representative docs: {representative_docs}
Label:

A.5.2 Radicalization Topic Labeling

You are analyzing a topic derived from the r/QAnonCasualties subreddit, where individuals share personal stories about their relatives' involvement in the QAnon conspiracy. These narratives often include emotional experiences, opinions, and real-world context.

Topic Keywords: {representation}
Representative Documents: {representative_docs}

Please provide a concise and descriptive topic label depicting the main theme of the above content. Be specific and don't infer any additional information that is not stated in the text. Be action-oriented and focus on the object of narration. Be concise, straightforward and no need to mention QAnon context in the label.

If the topic is unclear or contains multiple equally prominent themes, respond with:
Unsure - <Reason>

Format: <Concise Label>

A.5.3 Control Variable Extraction

For relationship:

You are a helpful assistant analyzing Reddit posts from the subreddit r/QAnonCasualties. Your task is to identify the relationship of the person they are describing (the object of narration, or OON, who radicalized) to the narrator (the person writing the post), if such a relationship exists.

If the narrator is telling a personal story about someone they know who became radicalized by QAnon, output the precise relationship label (e.g., "mother", "father", "brother", "sister", "wife", "friend").

If the post is not about a specific person known to the narrator (e.g., general discussion, support, resource sharing), or if you are not sure, return: "N/A".

If the post talks about a stranger or a person not clearly connected to the narrator, return: "stranger".

If the post talks about the radicalization/recovery story of the narrator themselves, return: "self".

Table A.4: Hierarchical structure of the Phase 3 (*Post-Radicalization*) topic consolidation. The rightmost column shows examples of the original, fine-grained topics that were grouped into the Sub-Categories and Categories (shown in the main Sankey) on the left.

Behavioral Transformation (External)	Social Deterioration and Relationship Strain	Pushback Against Boundaries Communication Tension and Breakdown Social Isolation and Limited Friendships Family Estrangement and Disconnection
	Interpersonal Harm and Moral Disregard	Unfaithful Apology Increased Aggression and Hostility Toward Others Loss of Compassion and Empathy Framing Others as Indoctrinated Emotional and Physical Abuse
	Alternative Media Use and Information Exposure	Distrust in Mainstream Media Joining and Using Alternative Social Media Platforms (Parler, Gab) Using Telegram for QAnon Information Excessive Screen Time and Online Content Consumption Disseminating QAnon Content and Persuading Others
	Lifestyle Changes Driven by Conspiratorial Beliefs	Increased Gun Acquisition Precious Metals Investment Prepping and Hoarding Supplies Homeschooling Children Employment Instability and Career Disruption
	Psychological and Emotional Shifts	Manifestations of Intense Hatred Collective Narcissism Inflated Self-Perception as Intellectual Superiority Willful Ignorance Personality Changes
	Radical Ideological Alignment and Intolerance	Racism Attitudes Negative Views on LGBTQ+ Communities Fascism Accusations and Rhetoric Deflecting to Race-Based Movements
	Public Health Mistrust and Resistance	Vaccine Refusal and Hesitancy Denying COVID-19 Reality Resistance to Mask-Wearing Resistance to Therapy and Counseling Distrust in Medical Professionals
	Epistemic Rigidity and Distrust in Expertise	Rejecting Verified Sources and Clinging to Unverified Information Distrust in Science Using "Science" to Justify Conspiracy Beliefs Insistence on Personal Research as Evidence Uncritical Acceptance of Beliefs and Theories Lack of Critical Thinking
Beliefs (Internal)	Child Harm and Elite Abuse Narratives	Child Sex Trafficking and Satanic Sacrifice Pizzagate Epstein-related Beliefs Disney "Hidden Imagery"
	Apocalyptic and Eschatological Narratives	The Cabal and "The Storm" Apocalyptic and Predictive Programming "10 Days of Darkness" "Trust the Plan" Lizard People Religious End Times Alien
	Pro-Trump	Trump's Return and January 6th Insurrection Trump as a Religious Figure MAGA Biden Presidency Denial Negative Reactions to Kamala Harris and Political Opposition JFK Jr. Being Alive
	Anti-Government and Institutional Distrust	Deep State New World Order Election Fraud
	Medical Mistrust	Drinking Bleach as a Cure Ivermectin Use in Treating Illness Pandemic Misinformation Dr. Fauci Creating COVID
	Science and Technology Mistrust	Climate Change is not Real 5G Being Dangerous Fake Moon Landing Flat Earth
	War-Inducing Rhetoric	Antisemitic Beliefs and Holocaust Denial Russia-Ukraine War Propaganda Invoking War, Freedom, and the Constitution
	Defamatory Conspiracy Targeting Public Figures	Bill Gates Michelle Obama Being a Man
	Policy Defiance	Tax Evasion Abortion Opposition
	Anti-Communism and Geopolitical Distrust	China and Global Influence Concerns Criticism towards Communism and Socialism

```
If multiple relationships are discussed, return
the closest and most central one,
prioritizing immediate family or partner.
```

```
Input:
Title: {title}
Body: {selftext}
```

```
Output:
Relationship:
```

For age group:

```
You are a helpful assistant analyzing Reddit
posts from the subreddit r/QAnonCasualties.
Your task is to estimate the narrator's age
group, which is either "adolescence" (13-20)
or "adult" (21+).
If unsure, return: "N/A".
```

```
Input: "{text}"
```

```
Output:
```

For gender:

```
You are a helpful assistant analyzing Reddit
posts from the subreddit r/QAnonCasualties.
Your task is to estimate the narrator's
gender, which is either "f" (female) or "m"
(male).
If unsure, return: "N/A".
```

```
Input: "{text}"
```

```
Output:
```

A.5.4 Emotion Classification

```
You are an emotionally-intelligent and
empathetic agent.
You will be given a piece of text derived from
posts in the subreddit r/QAnonCasualties,
which are personal narratives describing one
's experiences with a loved one who has been
radicalized.
You must identify the emotions explicitly stated
or clearly implied by the writer about
their own feelings.
Only consider the writer's emotions - ignore the
emotions of other people mentioned.
If the emotion is expressed at least once in the
text, output 1; otherwise, output 0.
Do not infer emotions based solely on the topic
or context unless the writer states or
clearly implies them.
Possible emotions: Anger, Disgust, Sadness, Fear,
Surprise.
```

```
Text: "{text}"
```

```
Please provide outputs for these emotions in
JSON format:
```

```
{{
  "anger": 0 or 1,
  "fear": 0 or 1,
  "sadness": 0 or 1,
  "disgust": 0 or 1,
  "surprise": 0 or 1,
```

```
}}`
```

A.6 Exploring Methods for Identifying Personas

Our trajectory analysis, visualized in the main paper's Sankey diagram (Figure 1), reveals a complex system with a high degree of interconnectedness between thematic traits. We observe no clear, hard boundaries that would suggest mutually exclusive clusters of radicalization journeys. This data structure strongly indicates that hard clustering methods such as k-means are ill-suited, as they would force each individual's profile into a single, exclusive archetype, failing to capture the blended nature of their experiences. This observation also aligns with the theoretical understanding of radicalization as a complex and multifaceted phenomenon where different ideological and personal pathways often converge. Therefore, we seek a soft clustering method that allows for mixed membership. This choice narrows our evaluation to two primary candidates: NMF, which provides an additive decomposition, and LDA, which offers a probabilistic mixture.

To empirically compare LDA and NMF, we train models for both methods with the number of personas (k) ranging from 4 to 15. We compare the models using the C_v topic coherence score. As shown in Figure A.2, the LDA models generally achieves higher coherence scores than the NMF ones. Manual inspection on a selected few samples also verifies the superior interpretability of the LDA approach. Based on this convergent evidence, we select the LDA-inspired model for our final analysis.

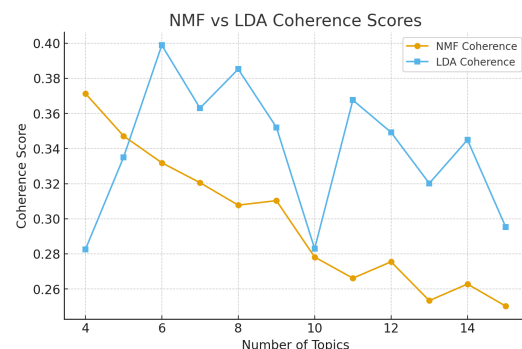


Figure A.2: NMF vs LDA coherence score.

A.7 Selection of Persona Number

As seen in the coherence score plot (Figure A.2), the score for LDA peaks at $k = 6$. While this provides a strong quantitative signal, the coherence

scores for $k = 8$ and $k = 11$ are also high, necessitating a qualitative review to select the most meaningful and non-redundant set of personas.

We then conduct a qualitative review of the models with the highest coherence scores ($k = 6, 8$, and 11), examining the top traits, sub-Sankey diagrams, and representative narratives for each persona. Our review reveals that models with higher k values often produces “fractured” personas, in which a coherent archetype is unnecessarily split into multiple, redundant sub-personas. For example, as illustrated in Figure A.3, the $k = 8$ model generates two personas centered on the Trump Election trigger. One emphasizes radical ideologies and anti-authority themes, and the other medical distrust, yet both are minor variations of the same overarching Political Extremist archetype already present in the $k = 6$ model. This redundancy would dilute clarity without adding new insight. In contrast, the $k = 6$ model produces the most distinct and coherent personas, with each representing a unique radicalization pathway and minimal thematic overlap. Based on this convergent evidence—the peak coherence score at $k = 6$ and its superior qualitative distinctiveness and non-redundancy—we select $k = 6$ for the final analysis.



Figure A.3: An example of a “fractured” persona from the $k = 8$ LDA model.

A.7.1 Persona Stability Analysis

We conducted a stability analysis of the final $k = 6$ persona model by re-fitting LDA five times with different random seeds (42, 123, 456, 789, 999). Since persona ordering is arbitrary across independent LDA runs, we first aligned personas across runs using the Hungarian algorithm on topic-profile

Table A.5: Per-persona stability: average cosine similarity (topic-profile β) between matched personas across five LDA re-fits with different seeds. Persona numbers correspond to Table 2.

Persona	Cosine Sim. (β)	SD
P1	0.712	0.167
P2	0.612	0.164
P3	0.409	0.067
P4	0.591	0.181
P5	0.636	0.137
P6	0.956	0.034
Mean	0.639	0.219

(β) similarity, then evaluated stability of both the persona-trait distributions (β) and document-persona allocations (θ) using this fixed alignment. We note that cosine similarity on 50-dimensional trait distributions is inherently sensitive to run-to-run variation in the low-probability tail of the distribution.

Across the five runs, matched personas showed an average cosine similarity of 0.639 (SD = 0.219) for topic-profile (β) distributions and 0.667 (SD = 0.139) for document-topic (θ) allocations. The coherence score that determines $k = 6$ varied minimally across seeds (SE = 0.015).

Stability was not uniform across the six personas, as shown in Table A.5. Persona 6 (Pandemic-Triggered Skeptic) was the most stable by a wide margin (cosine similarity = 0.956, SD = 0.034), while Persona 3 (Social Media Spiral) showed the lowest average similarity (0.409)—though notably, its low standard deviation (0.067) indicates this is a *consistently* lower-cosine match across seeds rather than an erratic one. The remaining four personas fall in a moderate mid-range (0.591–0.712); of these, Persona 1 (Health-Triggered Conspiracy Theorist) has the second-highest variability (SD = 0.167), suggesting it is not fully immune to run-to-run instability despite its higher mean similarity. We report this per-persona variation transparently and note it as a limitation: results involving lower-stability personas, particularly Persona 3, should be interpreted with correspondingly more caution.

A.8 Persona Representation

To interpret the qualitative substance of each of the six data-driven personas, we first identified and manually analyzed the top three narratives with the highest probability score for that archetype. Table A.6 details the distilled insights from that

analysis. For clarity, we present these insights as narrative composites. These composites are not direct quotes from a single post; they are synthesized stories that weave together the key themes, events, and emotional arcs from the most representative narratives. This technique, known as composite narrative construction (Willis, 2019), allows us to provide a single, rich, and illustrative example for each persona, complementing the concise descriptions and data visualizations in the main paper.

A.9 Persona Statistics

We also provide additional descriptive statistics for the six radicalization personas discovered in our analysis. These statistics offer further insight into their overall prevalence, their composition within individual narratives, and their evolution over time. Figure A.4 shows the average probability for each of the six personas across all analyzed profiles.

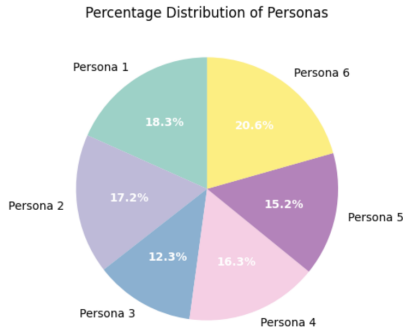


Figure A.4: Overall Distribution of Personas.

The distribution of occupied personas is depicted in Figure A.5. A profile is considered having an occupied persona X if the probability assigned to persona X is more than $1/6 \approx 0.167$. Since there are six possible personas in total, a single profile may be associated with more than one persona, anywhere from one up to five at the same time.

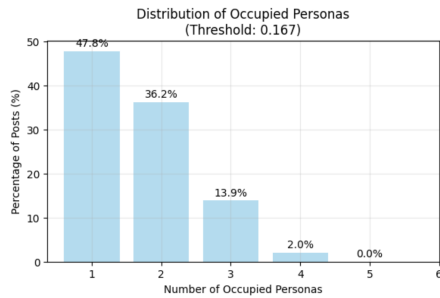


Figure A.5: Distribution of Occupied Personas.

Finally, we also illustrate the temporal dynamics of these archetypes over the time period of the dataset (2019-2024) in Figure A.6:

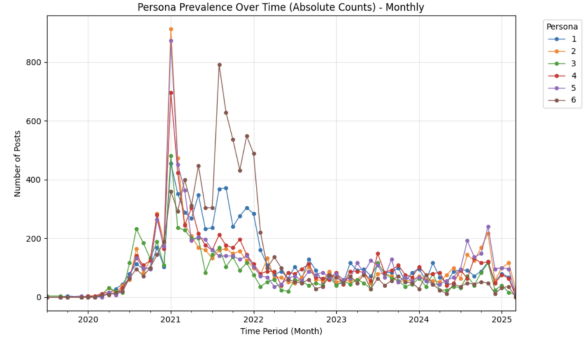


Figure A.6: Evolution of Personas over Time.

A.10 Validation of the LLM Emotion Annotator

A.10.1 Method Comparison

To validate our choice of LLM-assisted annotation, we compare three approaches on a 100-post sample independently annotated by two authors (annotators A and B): our GPT-based method (gpt-4o-mini), a lexicon-based approach (NRC Emotion Lexicon), and a fine-tuned classifier (GoEmotions; roberta-base-go_emotions)². We report Cohen's κ between each method and each human annotator in Table A.7. GPT-based annotation substantially outperforms both baselines across all emotions and annotators, achieving mean $\kappa = 0.71$ and 0.50 against annotators A and B, respectively. NRC and GoEmotions show near-zero agreement with human judgment (mean $\kappa \approx 0.00 - 0.05$), suggesting they are poorly suited to the long-form, multi-actor nature of these narratives. These results confirm that GPT-based annotation most closely approximates human judgment and is the most appropriate method for this task.

A.10.2 Annotator Agreement

We provide a fuller validation of the chosen GPT-based method, reporting full confusion matrices alongside precision, recall, F1, and Cohen's κ for each emotion against each human rater (Table A.8). The annotator achieves high raw accuracy and F1 (mean F1 = 0.91 vs. Rater A, 0.85 vs. Rater B) across all four emotions.

Error profiling shows the annotator is high-recall (92–100%) and its errors are dominated by false positives (24 of 27 disagreement cases), meaning it rarely misses a genuine emotional expression but sometimes over-attributes one; Appendix A.10.3 provides a qualitative taxonomy of these cases.

²https://huggingface.co/SamLowe/roberta-base-go_emotions

Table A.6: The narrative composites representing six radicalization personas.

Persona	Narrative Composite
The Health-Triggered Conspiracy Theorist	He used to be such a bright, intelligent, and sensible person. We were close friends since grade school, and he had a promising career as an engineer. But over the last couple of years, things took a dark turn. It started with some underlying health issues and escalated as he began mixing substances he was buying online. He started having these terrifying psychotic episodes, ranting about Q-adjacent theories and believing he was being watched. It got so bad his roommates had to call the police to have him hospitalized. He lied to me about it, claiming the doctors were performing experiments on him. He refuses to get professional help because he's convinced all doctors are evil and will just "inject him with shit." He eventually lost his job, and his spouse and I are watching him deteriorate. Now he spends his days in front of the computer, convinced the online community he interacts with knows more than anyone, even ranting at his wife—who is a nurse—about miracle cures he can't even pronounce. He used to be my best friend, and now he's just a shell of a human.
The Political Extremist	My relationship with a close family member has been completely destroyed by their political identity. They don't just support Trump; they worship him like a deity. Their entire worldview is based on hatred, fear, and a perceived sense of victimization, a constant us versus them mentality that echoes the darkest moments in history. This isn't hyperbole; the rhetoric is genuinely terrifying. They are trapped in a right-wing media bubble. I've tried to reason with them, carefully vetting my own sources to present non-partisan facts, but anything that contradicts their narrative is dismissed as "MSM Propaganda." Their radicalization has had severe real-world consequences: they lost their job as a teacher over their social media posts, were present at the January 6th riot, and have alienated everyone who tries to get close to them. We even attempted an intervention to show them how these beliefs were ruining their life, but it ended with them getting angry and trying to start a fight. I've had to end the relationship because I can't be close to someone whose values have become so full of hate.
The Social Media Spiral	My friend is very dear to me, but over the last few years, she's been consumed by conspiracy-minded YouTubers and alt-right talking points. She's completely Q-adjacent, believing everything they tell her to fear and be angry about. She thinks she's thinking for herself by using alternative media sources, but she's just being misled by paranoid video bloggers. Her entire reality is now shaped by these online rabbit holes. She's obsessed with wild theories about public figures being involved in child harm and elite abuse, things that started on fringe message boards like 8chan. It's a gigantic, ridiculous lie, but she takes it as gospel. I feel terrible for the people she and thousands of others baselessly accuse online of running satanic cults and using children's blood for anti-aging cream. When she visits, I want to see her, but I dread hearing about the latest thing some influencer said. It's exhausting, like they've co-opted the idea of breaking free from mental control only to submit to another, even uglier, form of it.
The Religious Apocalypticist	My wife used to be my partner, but now she's lost in a world where conspiracy theories and religious prophecy are one and the same. It started years ago with commentators like Alex Jones and spiraled into QAnon around 2018. She now believes in everything: the deep state, lizard people, adrenochrome, all of it. For her, and many in her conservative Christian circles, QAnon isn't just a conspiracy; it's a confirmation that the end times are here. Every major event is filtered through this lens. The COVID vaccine was the "mark of the beast," a tool for population control created by Bill Gates. She was convinced something huge would happen on January 6th to keep Trump in power. While she no longer believes "the storm is coming" after so many failed predictions, the underlying conspiratorial worldview remains. As a fellow Christian, I know that no one knows the day or the hour, but for her, these theories have become a self-reinforcing prophecy. She has cut off family and friends who don't agree, and I feel like I'm living with a stranger.
The Conservative Identity Protector	My parents used to be people I considered good and loving. But since 2016, my mother has transformed into someone I don't recognize, consumed by a need to protect her conservative identity from a world she sees as a threat. She is a Black woman who has become obsessed with right-wing talk radio and fringe news, and her mind is deteriorating as a result. She is terrified that our kids are being "transed," sees an "invasion at the border," and is constantly making snide remarks about anyone who doesn't share her views. Our family is suffering. Her constant political ramblings and racist, bigoted comments have created a screaming hell in our own house. As her gay son, happily married and trying to start a family, it's devastating to know she votes for people who want to strip away my rights. After the last election, she texted me a series of cheerful emojis, asking how I was "holding up"—not out of concern, but to bait me into an argument and gloat. She wants to "own the libs" more than she wants a relationship with her own son. It's exhausting and I am so done trying to talk sense into her.
The Pandemic-Triggered Skeptic	My mom has always leaned towards alternative medicine, but the COVID-19 pandemic sent her over the edge. We managed to get through the first few years by tiptoeing around the subject, but now it's unavoidable. It started with soft questions about how many vaccines I'd gotten, and now it's full-blown propaganda. She's convinced the vaccines are a form of mass genocide designed to alter our DNA and make us infertile. She calls herself a truth seeker and spends her time on platforms like Telegram, deep in rabbit holes about satanic abuse and globalist plots. She rejects the term "conspiracy theory" and proudly states she'd rather lose her job than get vaccinated. I made the mistake of checking out a YouTube channel she recommended, and it was pure fear-mongering and right-wing propaganda. We used to be thick as thieves, but now I feel like I'm mourning the person she once was. Her pre-existing skepticism of mainstream medicine was the gateway, and the pandemic pushed her through it into a much darker world.

We note one apparent inconsistency: κ for sadness is low against Rater B (0.260) despite high raw agreement (90%) and F1 (0.946). We attribute this to the prevalence paradox (Feinstein and Cicchetti, 1990)—because sadness is expressed in 96% of the validation sample, chance agreement is extremely high ($P_e \approx 0.865$), which mechanically depresses κ even when raw agreement and classifier utility remain strong. We therefore report raw accuracy and F1 alongside κ throughout, as we take the former to better reflect classifier utility in this skewed

regime.

A.10.3 Qualitative Error Analysis

We conducted a qualitative analysis of the 27 cases where the LLM disagreed with both human annotators (who were themselves in agreement), identifying four recurring error categories:

- **Source misattribution (9):** the model assigns an emotion expressed by a quoted third party (e.g., a family member's own words within the post) to the narrator, rather than tracking

Table A.7: Cohen’s κ between annotation methods and human annotators on a 100-post sample.

Method	Human	Anger	Fear	Sadness	Disgust	Mean
GPT	A	0.583	0.779	0.658	0.820	0.710
	B	0.464	0.702	0.260	0.575	0.500
NRC	A	-0.020	0.000	-0.016	0.043	0.002
	B	0.019	0.000	0.138	0.059	0.054
GoEmotions	A	0.000	0.077	0.025	0.000	0.026
	B	0.000	0.090	0.078	0.000	0.042

Table A.8: Full agreement metrics between the LLM annotator and each of two human raters, per emotion, on the 100-post validation sample.

Human	Emotion	Prec.	Rec.	F1	Acc.	κ
A	Anger	0.700	1.000	0.824	79.0%	0.583
	Fear	0.857	0.941	0.897	89.0%	0.779
	Sadness	0.980	1.000	0.990	98.0%	0.658
	Disgust	0.906	0.923	0.914	91.0%	0.820
B	Anger	0.657	0.939	0.773	73.0%	0.464
	Fear	0.786	0.936	0.854	85.0%	0.702
	Sadness	0.898	1.000	0.946	90.0%	0.260
	Disgust	0.868	0.767	0.814	79.0%	0.575

the narrator’s actual stance toward that content. For instance, one post quotes a father’s angry reply accusing his daughter of unfairly calling him racist; the model flagged the post as expressing disgust, even though the disgust belonged to the quoted father, while the narrator’s own reaction throughout the post was one of sadness at the relationship’s breakdown.

- **Lexical bleed from topic content (8):** emotion-adjacent vocabulary (fear/threat language, violent content) triggers a positive prediction independent of the narrator’s actual emotional register, which is often closer to grief, exhaustion, or contempt. For example, a post describing a mother who wished serious illness on people who voted differently, and who threatened family members over politics, was flagged for fear despite the narrator’s own tone reading as weary contempt rather than fear.
- **Register/rhetorical confusion (6):** sarcasm, dismissive idiom, and rhetorical questions carry surface intensity markers that are misread as the literal emotion, when the narrator’s stated or contextual affect is different (and sometimes named explicitly elsewhere in the same post). For example, a narrator closed a post with a sarcastic line blaming several public figures and institutions for estranging them from a family member; the model read this as anger, even though the narrator explicitly

named their emotion as sadness a few lines earlier.

- **Implicit/under-lexicalized emotion under-weighted (4):** when an emotion is conveyed through consequence, somatic description, or moral disapproval rather than explicit affect vocabulary, the model tends to miss it, suggesting a reliance on explicit lexical cues over pragmatic inference. For instance, a health-care worker described repeatedly having to intubate COVID patients and notify families of deaths while trying to protect her infant son from unvaccinated relatives; despite the clear underlying fear, the model did not detect this emotion, likely because it was conveyed through professional and clinical detail rather than explicit fear-related wording.

A.11 Regex for Control Variable Extraction

The high-precision regular expression patterns used to create the ground-truth dataset for validating our LLM-based control variable extraction are inspired by prior work on demographic inference from social media text (Tigunova et al., 2020), refining their patterns based on common phrasing observed in our dataset. Table A.9 outlines the general syntactic patterns we search for, while Table A.10 lists the specific keywords and phrases used as indicators within those patterns.

After identifying these variables, *gender* is coded as a binary variable (male or female), and *age group* is grouped by adolescence (13-20) and adult (21+) for the demonstrated usage and performance in social media analysis (Cesare et al., 2017) and the distribution of user’s age shown in Figure A.7

Table A.9: Syntactic patterns for labeling narrator age and gender. Placeholders like ‘<...>’ are filled with indicators from Table A.10.

Attribute	Pattern(s)
gender	(I am I’m) a? ‘<gender indicators>’
age	(i) (I am I’m) ‘<number>’ years old (ii) I’m a ‘<number>’-year-old ‘<gender indicator>’ (iii) I am ‘<number>’ followed by punctuation/conjunction (iv) Compact Format: ‘<number><gender indicator>’

Table A.10: Word and phrase indicators used within the patterns from Table A.9.

Attribute	Value	Word/Phrase Indicators
gender	female	woman, female, girl, lady, wife, mother, sister
	male	man, male, boy, guy, husband, father, brother

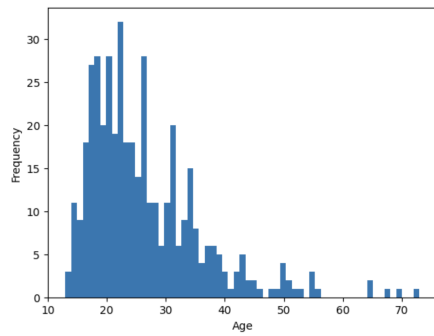


Figure A.7: Distribution of ages of authors.

A.12 Full Emotion Regression Coefficient Table

The full regression coefficients for RQ2 are shown in Table A.11. The categorical control variables are coded using effect coding, where we interpret their effect from the overall mean, given there is no natural baseline for choosing a reference category (e.g., among different family relationships).

Table A.11: Regression Results (Odds Ratios and p-values, rounded to 3 decimals). Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	OR (anger)	p (anger)	OR (disgust)	p (disgust)	OR (fear)	p (fear)	OR (sadness)	p (sadness)
Intercept	2.305***	0.000	1.509***	0.000	1.656***	0.000	7.933***	0.000
C(relationship_cat)[extended_family]	1.038	0.075	1.231	0.058	0.988	0.912	0.826	0.255
C(relationship_cat)[friend]	0.615***	0.000	0.995	0.962	0.665***	0.000	0.949	0.075
C(relationship_cat)[parent]	1.458***	0.000	1.090	0.092	1.151**	0.006	1.331**	0.001
C(relationship_cat)[partner]	0.733***	0.000	0.664***	0.000	1.355***	0.000	0.691***	0.000
C(age_group)[adolescence]	0.878***	0.001	0.922*	0.016	1.174***	0.000	0.776***	0.000
C(gender)[f]	0.918*	0.038	0.898**	0.000	0.980	0.059	1.098	0.113
balance_1	0.927**	0.041	0.870***	0.000	1.045	0.069	0.928	0.067
balance_2	1.036	0.230	0.878***	0.000	1.087**	0.023	1.109*	0.031
balance_3	1.024	0.343	0.973	0.214	0.897***	0.000	1.108*	0.012
balance_4	1.108**	0.050	1.086***	0.000	0.936***	0.009	1.025	0.057
balance_5	0.999	0.977	0.898***	0.000	0.986	0.057	0.989	0.808