Detecting Misconception Surges in Opioid Policy: A Zero-Shot LLM Analysis of British Columbia's Decriminalization Pilot

Sepehr Harfi Moridani
University of Toronto
Marwa Sharifi
University of Toronto
Chan Pang Yang
University of Toronto
Konrad Samsel
University of Toronto
Christoffer Dharma
chris.dharma@mail.utoronto.ca
University of Toronto
Mohammad Noaeen
University of Toronto
Zahra Shakeri
University of Toronto

Research Article

Keywords: Decriminalization, Drug policy, Social listening, policy misperceptions, Harm reduction communication, Large language models

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8897658/v1

Additional Declarations: No competing interests reported.

Detecting Misconception Surges in Opioid Policy: A Zero-Shot LLM Analysis of British Columbia's Decriminalization Pilot

Sepehr Harfi Moridani[1,+], Marwa Sharifi[1,+], Chan Pang Yang[1], Konrad Samsel[1], Christoffer Dharma[1], Mohammad Noaeen[1], and Zahra Shakeri[1,2,3*]

[1]Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[2]Faculty of Information, University of Toronto, Toronto, Canada
[3]Schwartz Reisman Institute, University of Toronto, Toronto, Canada
[*]Corresponding author, chris.dharma@mail.utoronto.ca
[+]These authors contributed equally to this work

Abstract

Background: In January 2023, British Columbia (BC) launched Canada's first drug decriminalization pilot in response to persistently high opioid-related deaths. This initiative removed criminal penalties for possession of up to 2.5 grams of certain illicit drugs to reduce stigma and increase service use. Social media platforms such as Reddit provide a real-time window into public perceptions and misconceptions about such policies.

Methods: We collected 22,131 comments posted between January 2023 and October 2024 from subreddits discussing decriminalization in British Columbia. A zero-shot large language model applied a multi-label taxonomy with four misconception classes and one no-claim class. Expert adjudication of 547 comments with model disagreement supported selection of the primary classifier for analysis. We tracked monthly misconception prevalence and used subreddit-week reply premiums and lattice clustering to detect engagement bursts. Negative binomial models quantified reply differences across periods, regions, and comment-level harm reduction references.

Results: GPT-4o labeled 4,473 comments as containing at least one misconception, and policy-target misinterpretation and impact claims dominated through early 2024. After the May 2024 amendment that restricted possession in public spaces, *'Legal-Status Confusion'* and enforcement misconceptions increased and peaked in June 2024. These comments received an 11% reply premium, with an incidence rate ratio of 1.11 (95% CI 1.02-1.21), and the premium increased after May 2024 (1.18, 1.16-1.21). Reply premiums clustered into four bursts from February to July 2024, and each burst lasted 2 to 5 weeks across 2 to 5 subreddits. Harm-reduction terms predicted higher replies, with an incidence rate ratio of 1.12 (95% CI 1.04-1.20), and the interaction ratio was 0.81 (0.66-0.98). Community-level harm-reduction visibility did not predict week-to-week variation in reply premiums for either misconception class.

Conclusions: Public discussion of BC's decriminalization policy revealed persistent confusion, and legal and enforcement confusion intensified around the May 2024 revision. Burst detection on social platforms can identify when and where confusion concentrates, which supports timely public clarification and service signposting. During future amendments, concise messages about permissible locations and remaining police authority may reduce avoidable enforcement contact and misinformation.

Keywords: Decriminalization, Drug policy, Social listening, policy misperceptions, Harm reduction communication, Large language models

1 Introduction

The opioid crisis is a major public health challenge in Canada and has caused over 50,000 deaths since 2016.[1] In 2024 alone, 7,146 opioid related deaths were reported, an average of 20 deaths per day.[1] Opioids include prescription analgesics such as morphine and oxycodone as well as illicit substances like fentanyl and heroin.[2,3] The route of administration, frequency of use, and individual tolerance levels all shape the risk of harm.[4,5] Co-occurring mental health conditions and limited access to harm reduction services can heighten the likelihood of dependence or overdose.[6] The financial toll is high, with the total economic burden of opioid use in Canada estimated at $7.1 billion annually, including hundreds of millions spent directly on healthcare and emergency interventions.[7,8,9] In 2020, substance use cost Canada's healthcare system $13.4 billion, with opioid-related healthcare costs alone accounting for $519 million.[9] Despite significant investments in healthcare and prevention, opioid toxicity deaths remain high, with the highest reported cases in British Columbia (BC), Alberta (AB), and Ontario (ON).[1] Many factors contribute to the ongoing opioid crisis and poor patient outcomes, with stigma being a significant one. Stigma can discourage individuals from seeking help due to fear of judgment, and often results in lower quality care when accessing services.[10] A qualitative study conducted in BC found that people who use drugs usually face discrimination in healthcare settings, are denied pain medications, and are

Table 1: Characteristics Reddit posts related to decriminalization in British Columbia by subreddit. Rightmost column shows a miniature filled line of monthly post counts with a shared vertical scale across rows; all densities cover 2023-01-31 to 2024-10-31.

Subreddit	Posts	Total replies	Mean replies (SD)	Total upvotes	Median upvotes [IQR]	Date range	Density (monthly posts)
r/canada	9603	6815	0.7 (1.0)	52515	2 [0–4]	2023-02-22 to 2024-09-08	
r/vancouver	2978	2045	0.7 (1.0)	28018	2 [0–7]	2023-05-14 to 2024-10-20	
r/britishcolumbia	2259	1450	0.6 (1.0)	14884	2 [1–4]	2023-01-31 to 2024-08-04	
r/Canada_sub	1711	1144	0.7 (0.9)	3090	1 [1–2]	2023-07-19 to 2024-07-27	
r/VictoriaBC	1336	938	0.7 (1.0)	5792	2 [1–4]	2023-08-13 to 2024-08-30	
r/CanadaPolitics	1040	858	0.8 (1.0)	5394	2 [1–6]	2024-04-22 to 2024-07-30	
r/onguardforthee	1016	607	0.6 (0.9)	6934	2 [1–6]	2023-02-22 to 2024-10-08	
r/ontario	872	626	0.7 (1.1)	4403	2 [1–4]	2024-03-30 to 2024-06-07	
r/toronto	551	386	0.7 (0.9)	2712	2 [1–5]	2024-05-01 to 2024-05-20	
r/BCpolitics	301	222	0.7 (0.7)	580	1 [0–2]	2023-05-18 to 2024-09-18	
r/saskatoon	276	211	0.8 (1.2)	920	1 [1–4]	2024-04-29 to 2024-05-22	
r/saskatchewan	127	94	0.7 (0.8)	559	2 [1–4]	2024-04-29 to 2024-05-01	
r/UBC	55	29	0.5 (0.6)	354	3 [1–6]	2023-02-01 to 2023-02-03	
r/NiceVancouver	6	0	0.0 (0.0)	8	1 [1–2]	2023-08-14 to 2023-08-15	
All subreddits	**22,131**	**15,425**	**0.7 (1.0)**	**126,163**	**2 [1–5]**	**2023-01-31 to 2024-10-20**	

labeled as 'drug-seeking'. These experiences can lead to self-medication with illicit substances and increase overdose risk.[11] In response, several regions worldwide have considered or implemented decriminalization to reduce stigma and encourage treatment.[12,13] Oregon initially decriminalized small amounts of hard drugs in 2023, but reversed course in September 2024, citing policy and implementation hurdles.[14] In Canada, BC became the first and only province to implement a decriminalization policy. On January 31, 2023, Health Canada granted BC a three-year exemption allowing adults to possess up to 2.5 grams of certain illicit drugs for personal use in designated locations.[15] The aim of the policy was to reduce stigma and encourage people to seek health and social services rather than penalizing them under the criminal justice system.[16]

Early evidence has also begun to document how British Columbia's decriminalization pilot and its May 2024 amendment have been experienced in frontline service settings. Using key-informant interviews with harm reduction and opioid agonist treatment providers across the province, Russell et al. (2025) reported few immediate operational changes following decriminalization, but emphasized persistent resource constraints and substantial uncertainty linked to limited policy communication and site-specific training, particularly around the May 2024 revision that re-criminalized public use and possession.[17] These implementation experiences suggest that decriminalization's public health impact depends not only on the legal change itself, but also on how clearly the policy is communicated and understood across stakeholders. Despite these early implementation insights, less is known about how the public understands, or misunderstands, policy details in real time and how confusion intensifies during policy changes, especially in large-scale and high-intensity online discussions.

Public opinion plays a key role in the policy's success, as it can influence how widely the policy is accepted and whether at-risk individuals feel safer seeking help. However, while public engagement is often beneficial, it can also lead to the spread of misinformation or confusion about the policy's details, which in turn may hinder its effectiveness.[18,19] Misinformation or misconceptions about what 'decriminalization' entails may spread quickly. For example, people may confuse decriminalization with full legalization[20,21] or misunderstand how the police are enforcing the new rules, which can lead to the rapid dissemination of false ideas.[17] Such misconceptions can undermine public trust and hinder effective implementation.

Social media platforms, such as Reddit, provide a large pool of timely and organic user perspectives on policy changes[22] Unlike traditional surveys, which can be limited by recall bias or social desirability effects,[23,24] Reddit discussions often include candid reflections on drug use, harm reduction, and policy outcomes.[22] Analyzing these discussions can help identify common points of confusion and concern in real time, which provides valuable feedback to policymakers and public health communicators.

In this study, we address three key questions. First, what are the prevalent misconceptions about BC's decriminalization policy expressed in online discussions, and how did the prominence of these themes change over time? Second, do comments containing misperceptions about the policy's legal status or enforcement attract unusually high levels of engagement, and if so, when do these intense discussions ('bursts') occur? Third, do these comments containing misperceptions have higher levels of engagement when there are more visible harm-reduction content in a given community? By answering these questions, we aim to explore *what* the public tends to misconceive about the policy, *when* these misunderstandings generate surges of engagement, and *how* public engagements in these misperceptions are potentially modified with increased harm reduction messaging. Clarifying these aspects can inform targeted communication strategies. For instance, knowing which specific misunderstandings to address and when to intervene (e.g., during intense online debates or after policy updates) can help public health officials focus their communication efforts where they are most needed.

2 Materials and Methods

2.1 Data Collection

We collected a comment-level Reddit corpus to study public discourse following the implementation of British Columbia's drug decriminalization policy (effective January 31, 2023). We queried Reddit for the case-insensitive phrase 'BC Decriminalization' and, for each matched thread, retrieved post metadata (e.g., title, URL, timestamp) and programmatically extracted all top-level comments and their nested replies. Each observation in the analytic dataset corresponds to a single comment and their replies with the following fields: raw comment text, number of upvotes, publication timestamp, subreddit name, and direct reply count. Personally identifiable information (e.g., usernames) was not collected at any stage of the study.

Table 1 summarizes the dataset by subreddit. For each community, we report the number of posts and comments, total replies, mean (SD) replies per comment, total upvotes, median upvotes with interquartile range (IQR), and the calendar coverage of ingested comments. To provide a compact picture of temporal activity within the table, the rightmost column shows a miniature density of monthly comment volume drawn on a *shared* vertical scale across rows.

As summarized in Table 1, the corpus contains 22,131 comments from January 2023 to October 2024, with activity concentrated in a few large communities: `r/canada` (9,603; ≈43% of all comments), `r/vancouver` (2,978; ≈13%), and `r/britishcolumbia` (2,259; ≈10%). Participation is limited but consistent across subreddits (i.e., mean replies per comment ≈0.7, SD 1.0) and upvotes are low to moderate (overall median 2, IQR 1–5). The inline density glyphs show a clear spike in early 2024 that peaks in `r/canada` and stays lower in BC-focused communities (e.g. `r/vancouver` and `r/britishcolumbia`). The dataset used in this study and the code to reproduce the pipeline are available for replication and further analysis.[25]

2.2 Policy Misconception Classification

We operationalized *policy misconception* as a checkable factual error about the intent, legal scope, enforcement practice, or alleged downstream impacts of BC's decriminalization exemption. We built a five-class taxonomy to balance policy relevance and labeling reliability. Four classes captured misconception types that recur in public debate and appear in prior work on decriminalization knowledge gaps, definitional confusion, and policy framing.[26,27,28,20] A fifth class captures comments that contain no verifiable factual claim and therefore cannot be evaluated as correct or incorrect. We kept the taxonomy small because we wanted consistent annotation at scale and clean downstream interpretation for a public health audience. Table 2 lists these labels, along with the operational definition we applied during labeling and an illustrative example for each class. We designed the classes to map onto concrete policy elements that are often misunderstood: the policy purpose (Label 1), the difference between decriminalization and legalization (Label 2), enforcement powers and limits (Label 3), and causal claims about crime or disorder (Label 4). We used Label 5 only when the comment contained no checkable factual statement about the policy.

Since many comments combine multiple factual errors in one post (e.g., conflating legal concepts (Label 2), misstating police powers (Label 3), or reporting crime growth (Label 4)), we used a multi-label format and allowed up to three labels per comment to keep outputs focused and interpretable. We used zero-shot multi-label classification, where the model assigns labels from Table 2 without task-specific training examples.[29,30] This choice matched our goal of measuring how general-purpose models interpret policy claims in public discussion. It also avoided a custom training set tied to one period's phrasing, which would reduce comparability across time and communities.[31] We instructed the model to prioritize explicit factual claims over tone or political stance. During parsing, we treated Label 5 as a mutually exclusive category and removed it whenever it appeared alongside any of Labels 1-4.

To label policy misconceptions at scale and maintain comparability across weeks and subreddits, we used the following fixed and taxonomy-based prompt with GPT-4o and Gemini-2.0-Flash.[32,33] The prompt returned only valid label IDs, separated checkable policy claims from rhetorical opinion, and kept the label boundaries stable under slang changes and event-driven spikes.

Policy Misconceptions Prompt

```
You are a public health policy analyst. The task is to
identify up to three factual error types in a Reddit comment
about British Columbia's drug decriminalization pilot. Select
labels from this list:
• 1. Policy-Target Misinterpretation: The comment
  misrepresents the purpose of the policy. For example,
  it claims the policy promotes drug use rather than access to
  health and social services.
• 2. Legal-Status Confusion: The comment conflates
  decriminalization with legalization. For example, it
  claims drugs are fully legal or permitted everywhere.
• 3. Enforcement Misconceptions: The comment misstates
  police or authority powers. For example, it claims police
  cannot intervene, confiscate drugs, or act on trafficking.
• 4. Impact on Crime or Society: The comment attributes
  crime, overdose deaths, or disorder directly to the policy
  without credible support.
• 5. Unclear or no verifiable factual claim: Use this label
  only when no other label applies.
Apply these rules:
• The classification should focus on factual claims rather
  than tone or opinion.
• The output should include one to three labels. Use fewer
  labels when the evidence is weak.
• Label 5 applies to satire, unverifiable anecdotes, or
  comments without a checkable factual claim.
Output the label numbers only, separated by commas.
Comment: [Comment Body]
```

To quantify agreement under multi-label predictions, we used the mean normalized Hamming distance[34,35] between two label sources:

$$H(A,B) = \frac{1}{NC} \sum_{j=1}^{N} \sum_{c=1}^{C} \mathbf{1}\{y_{j,c}^{A} \neq y_{j,c}^{B}\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function, N is the number of

Table 2: **Policy misconception taxonomy used for multi-label classification.** Each comment can receive up to three labels. Examples are illustrative and paraphrased.

Label	Operational definition (what we label)	Illustrative example
1. Policy-Target Misinterpretation	The comment misstates the purpose of the exemption. It presents the policy as promoting or encouraging drug use, rather than reducing harms by shifting personal possession away from criminal penalties and toward health and social supports.	"They want everyone to start using drugs."
2. Legal-Status Confusion	The comment conflates decriminalization with legalization or regulation. It claims drugs are "legal", can be bought legally, or can be used anywhere without restriction.	"Drugs are legal now, so you can do them anywhere."
3. Enforcement Misconceptions	The comment misstates what police or other authorities can do. It claims enforcement has stopped entirely, that police cannot seize drugs, or that trafficking is permitted or ignored.	"Police cannot take drugs or stop dealers anymore."
4. Impact on Crime or Society	The comment asserts a direct, sole, or definitive causal link from the exemption to changes in crime, overdoses, or public disorder without credible support. We applied this label to causal claims framed as settled fact rather than as uncertainty or personal experience.	"Crime exploded because of decriminalization, nothing else."
5. Unclear or no verifiable factual claim	The comment contains no checkable factual claim about the policy. It is pure opinion, sarcasm, insult, or anecdote without a falsifiable statement about intent, scope, enforcement, or impacts. We used this label only when none of Labels 1-4 applied.	"This policy is a disaster".

comments, and C is the number of classes ($C = 5$). In the multi-label setting, each prediction is a binary vector over the misconception classes, with $y_{j,c} = 1$ when class c is present for comment j. The Hamming distance equals the number of class positions where the binary vectors differ. For example, consider $C = 5$ classes and one comment with two different predictions below:

- Model A predicts classes 1 and 5: $y^A = [1, 0, 0, 0, 1]$.
- Model B predicts classes 2 and 3: $y^B = [0, 1, 1, 0, 0]$.

These vectors differ at positions 1, 2, 3, and 5, so $H(y^A, y^B) = 4/5 = 0.8$ for this comment. We also tested LLaMA-3.2B as a lightweight open-source baseline during early pilot labeling.[36,37] Pairwise comparisons on the same comment set showed close agreement between GPT-4o and Gemini-2.0-Flash (0.0627). LLaMA-3.2B differed more from GPT-4o and Gemini-2.0-Flash (0.4331 and 0.4448), and pilot runs produced unstable label distributions and more taxonomy violations. These patterns led to the exclusion of LLaMA-3.2B from the main labeling pipeline.

To select the primary model and quantify alignment with expert judgment, we focused on comments where GPT-4o and Gemini-2.0-Flash disagreed. Across the full corpus, GPT-4o and Gemini-2.0-Flash disagreed on 3,644 comments in total. To keep manual review feasible and cover various disagreement cases, we selected a 15% sample ($n = 547$) from this set of disagreements. Two domain experts labeled this sample using the same five-class definitions in Table 2. We computed the mean Hamming distance between each model and each expert label vector and selected GPT-4o for the main analysis because it showed lower disagreement than Gemini-2.0-Flash on the manually reviewed sample.

2.3 Reply Burst Detection

Our second research question asked whether specific policy misconceptions triggered unusually high reply activity, and when these surges occurred. We focused on the *'Legal-Status Confusion'* and *'Enforcement Misconceptions'* categories because they directly concern the legal scope of the exemption and police powers. To compare replies to misconception comments against replies to other comments in the same subreddit and week, we measured burst-like engagement[38] at the subreddit-week level using a *reply premium*. If r_i denotes the direct reply count for comment i, we define the variance-stabilized outcome as

$$y_i = \log(1 + r_i).$$

For each subreddit s and calendar week w, we define $M_i = 1$ to indicate that comment i contains either a legal-status misconception or an enforcement misconception, and we define $M_i = 0$ otherwise. We compute the within-cell mean difference

$$\Delta_{s,w} = \bar{y}^{(1)}_{s,w} - \bar{y}^{(0)}_{s,w}, \qquad \bar{y}^{(m)}_{s,w} = \frac{1}{N^{(m)}_{s,w}} \sum_{i \in (s,w):\, M_i = m} y_i,$$

where $N^{(m)}_{s,w}$ is the number of comments in subreddit s and week w for which $M_i = m$. We then map this difference to a percentage scale:

$$\text{Premium}_{s,w} = 100 \times \left\{ \exp(\Delta_{s,w}) - 1 \right\}.$$

A positive value indicates higher reply engagement for misconception comments than for other comments within that same subreddit-week. If either group was absent in a subreddit-week (i.e., $N^{(1)}_{s,w} = 0$ or $N^{(0)}_{s,w} = 0$), $\text{Premium}_{s,w}$ was undefined and left blank in downstream visualizations. Since reply premiums vary across subreddits, we standardized each subreddit's weekly series. For each subreddit s, we computed

$$z_{s,w} = \frac{\text{Premium}_{s,w} - \overline{\text{Premium}_s}}{\text{SD}(\text{Premium}_{s,\cdot})},$$

where the mean and standard deviation were computed over weeks with defined premiums. We then defined sign-specific 'hot' thresholds using quantiles of the standardized distribution, where the positive cutoff z^+ was the qth quantile of $z_{s,w} : z_{s,w} > 0$ and the negative cutoff z^- was the qth quantile of

$-z_{s,w} : z_{s,w} < 0$, with $q = 0.85$. A subreddit-week was flagged as a candidate *premium* hot cell if

$$z_{s,w} \geq z^{+} \quad \text{or} \quad \text{Premium}_{s,w} \geq a,$$

and as a candidate *penalty* hot cell if

$$z_{s,w} \leq -z^{-} \quad \text{or} \quad \text{Premium}_{s,w} \leq -a,$$

with an absolute cutoff of $a = 10$ percentage points. This combined rule captures weeks that are extreme in relation to the typical variability of a subreddit and weeks that are practically large on the original percentage scale.

Spatiotemporal clustering into bursts. Next, we treated each flagged subreddit-week as a grid cell and formed clusters using 8-neighbor adjacency, where cells touch by an edge or a corner. Here, "spatial" refers to adjacency on the subreddit-week lattice (i.e., subreddit index × time), not geographic proximity. We computed connected components separately for premium and penalty cells. Each connected component was summarized by (i) the number of hot cells, (ii) its week span, (iii) its subreddit span, and (iv) a coherence measure defined as the fill ratio of hot cells within the component's axis-aligned bounding box:

$$\text{Fill} = \frac{N_{\text{cells}}}{(\text{weeks span}) \times (\text{subreddits span})}$$

We retained bursts that spanned at least 2 weeks, involved at least 2 subreddits, contained at least 4 hot cells, and had Fill ≥ 0.35. These requirements filter out isolated spikes and instead focus on surges of engagement that were sustained and involved multiple subreddits. To support interpretation, we visualized the weekly reply-premium surface as a subreddit-week heatmap and overlaid outlines of the retained burst components. The resulting burst maps are reported in the Results Section.

2.4 Harm reduction discourse and engagement with legal/enforcement misconceptions

To understand how harm-reduction discourse shapes engagement with legal status and enforcement misconceptions about BC's decriminalization policy, we analyzed how comments that misstate the scope of the exemption or police powers perform in terms of replies and how this engagement varies in weeks when harm-reduction content is more visible in a given community. To this end, we began with the comment-level Reddit corpus described above, with one row per comment and fields for raw text, subreddit, timestamp, number of direct replies, upvotes, and the multi-label misconception string. Because exports used different column names, we standardized headers and mapped each field to the first matching candidate name. Header names were set to lower case, white space was removed, and the selected columns were retained. Timestamps were parsed from several human readable formats and, when necessary, from epoch seconds or milliseconds. All times were converted to UTC, which supported derivation of week, hour-of-day, and weekday variables. Rows with a missing subreddit, missing time, missing reply count, or empty text were removed.

We defined BC focused communities as subreddits such as `r/vancouver`, `r/britishcolumbia`, `r/VictoriaBC`, `r/BCpolitics`, and `r/UBC`. These subreddits were coded as region *BC_local*, while all other subreddits were coded as region *National*. Indicator *post* equaled 1 on or after 1 May 2024 and 0 before, and comment text and label string were converted to lower case before pattern matching.

Harm reduction and misperception indicators. We identified explicit harm reduction content using a rule based approach.[39,40] After converting text to lower case, we searched for terms related to overdose response, drug checking, supervised consumption and injection sites, overdose prevention sites, syringe service programs, opioid agonist therapies, prescribed safer supply, and key provincial institutions. Comments that contained at least one of these terms were flagged as harm reduction. The existing label string encoded four misconception classes. For this section, we focused on comments that misrepresented the legal status of possession and use or the powers of police and other authorities. We therefore created a combined indicator for legal and enforcement misconceptions that was set to one when either of these classes was present and zero otherwise.

Comment-level model of reply counts. Let Y_i be the number of direct replies to comment i. We modelled Y_i with a negative binomial regression with a log link,

$$Y_i \mid X_i \sim \text{NB}(\mu_i, \theta), \qquad \log(\mu_i) = X_i^{\top}\beta.$$

The linear predictor included M_i, a policy-period indicator Post_i for comments on or after May 1, 2024, and a region indicator Region_i for BC-local versus national subreddits. We included all interaction terms among M_i, Post_i, and Region_i. We added the harm-reduction indicator HR_i and the interaction $M_i \times \text{HR}_i$. We adjusted for $\log(1 + \text{len}_i)$, upvotes, hour of day, weekday, and subreddit fixed effects. When the negative binomial fit was unstable, we fit a Poisson model with the same linear predictor.

We used cluster-robust standard errors[41] with clustering at the subreddit level and report incidence rate ratios with 95% confidence intervals for all focal terms. We also computed the predicted mean replies for combinations of M_i, HR_i, Post_i, and Region_i by averaging fitted values over the observed covariate distribution.

Community-level harm-reduction visibility and reply premium. To quantify harm-reduction visibility, we aggregated comments to subreddit-week units. For each subreddit s and week w, we computed

$$\text{HR\_share}_{s,w} = 100 \times \frac{\sum_{i \in (s,w)} \mathbf{1}\{\text{HR}_i = 1\}}{N_{s,w}},$$

where $N_{s,w}$ is the total number of comments in that subreddit-week. We then measured class-specific engagement with misconceptions using a reply premium on the log scale. Let r_i be the reply count for comment i and $y_i = \log(1 + r_i)$. For each class $c \in \{\text{Legal-status}, \text{Enforcement}\}$, we computed within each subreddit-week

$$\Delta_{s,w,c} = \bar{y}^{(1)}_{s,w,c} - \bar{y}^{(0)}_{s,w,c}, \qquad \text{PremiumPct}_{s,w,c} = 100 \times \Big\{\exp(\Delta_{s,w,c}) - 1\Big\},$$

where the superscripts (1) and (0) denote class-c misconception comments and all other comments in the same cell. At

Table 3: **Detailed Model Performance: Label Prevalence ($n, \%$), Density, and Inter-Model Distance.** *Avg. Labels* denotes Label Density (the average number of labels assigned per sample). *Mean Dist.* denotes the Mean Average Distance (the average of a model's pairwise Hamming distances to all other models). The distance is calculated using the multi-label Hamming formula: $H(A,B) = \frac{1}{NC}\sum_{j=1}^{N}\sum_{c=1}^{C}\mathbf{1}\{y_{j,c}^{A} \neq y_{j,c}^{B}\}$.

Model	L1	L2	L3	L4	L14 (Any)	L5	Avg. Labels	Mean Dist.
LLaMA-3.2B	11,387 (51.45%)	6,302 (28.48%)	7,872 (35.57%)	11,472 (51.84%)	16,441 (74.29%)	5,690 (25.71%)	1.93	0.4390
GPT-4o	1,238 (5.59%)	1,455 (6.57%)	1,268 (5.73%)	1,790 (8.09%)	4,473 (20.21%)	17,669 (79.84%)	1.06	0.2479
Gemini-2.0-Flash	530 (2.39%)	817 (3.69%)	1,058 (4.78%)	2,160 (9.76%)	3,512 (15.87%)	18,967 (85.70%)	1.06	0.2537

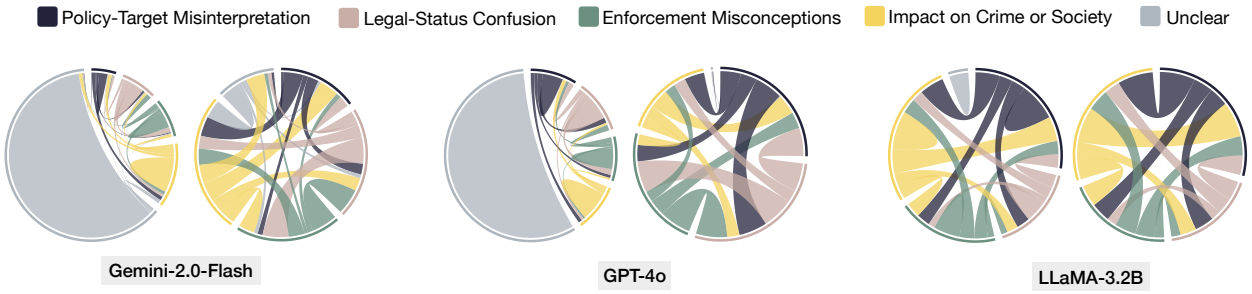

Figure 1: **Thematic Label Distribution and Co-occurrence Analysis.** Chord diagrams comparing the classification behavior of Gemini-2.0-Flash, GPT-4o, and LLAMA-3.2B across the five thematic categories: Policy-Target Misinterpretation, Legal-Status Confusion, Enforcement Misconceptions, Impact on Crime or Society, and Unclear. The diagrams on the left incorporate single-label assignments (self-loops) to represent the total prevalence of each category. The diagrams on the right depict only label co-occurrences and show the specific thematic intersections and overlapping logic identified by each model.

least three comments in each group were required to compute $\text{PremiumPct}_{s,w,c}$.

Finally, we tested whether harm-reduction visibility predicted reply premium variation across subreddit-weeks. A weighted linear regression was fit with $\text{PremiumPct}s, w, c$ as the outcome and $\text{HR}_s\text{hare}s, w$ as the main predictor. The model included region and class indicators, used robust standard errors, and weighted each observation by $N_{s,w}$.

3 Results

We compared model outputs using label prevalence, label density, and pairwise Hamming distance (Table 3). GPT-4o and Gemini-2.0-Flash placed most comments in Label 5, with 79.84% and 85.70% classified as unclear. Average Labels was 1.06 for both models, so most comments received one label and few received multiple labels. Label 5 acts as a null label that controls false positives when evidence for a misconception is weak. This distribution is expected for Reddit, where many posts express opinion, sarcasm, or anecdote without a checkable claim. However, LLaMA-3.2B produced a different profile, with 25.71% in Label 5 and Average Labels of 1.93. This pattern may result from weaker instruction following in a smaller model, which may treat the taxonomy as non-exclusive and over-assign labels. Its mean distance of 0.4390 also indicates that its outputs diverge from the other two models.

Figure 1 shows these differences by visualizing label co-occurrence patterns across the three models. Co-occurrence refers to comments that received more than one misconception label in a single prediction. The right-hand chords are sparse for GPT-4o and Gemini-2.0-Flash, which is expected given Average Labels near 1.0. Average Labels of 1.06 implies that at most about 6% of comments received a second label. In contrast, LLaMA-3.2B shows broad overlap among Labels 1-4, which aligns with Average Labels near 2.0. Average Labels of 1.93 implies that multi-label outputs were common and often included two labels. This overlap can occur when a model relies on policy keywords and assigns multiple misconceptions to a single stance statement. The post-processing rule that removes Label 5 when Labels 1-4 appear can also suppress Label 5 for LLaMA-3.2B. The dominant shared overlap for GPT-4o and Gemini-2.0-Flash links *'Legal-Status Confusion'* with *'Enforcement Misconceptions'*. This pairing tracks debates about what is legal under the exemption and what police can still do.

Within GPT-4o, most co-occurrence mass concentrates in a small set of pairings in the right panel (Figure 1). The strongest chord connects *'Legal-Status Confusion'* and *'Enforcement Misconceptions'* for multi-label outputs in GPT-4o. This pairing aligns with comments that treat decriminalization as full legality and then infer that police powers stopped. A prominent chord also links *'Policy-Target Misinterpretation'* with *'Impact on Crime or Society'*. This overlap corresponds to causal claims that treat the policy goal as promoting use and therefore increasing disorder. *'Legal-Status Confusion'* also co-occurs with *'Impact on Crime or Society'*, which captures posts that treat legality as a mechanism for crime or disorder. *'Policy-Target Misinterpretation'* co-occurs with *'Legal-Status Confusion'*, which indicates that some comments collapse intent and scope. Thin residual chords imply that GPT-4o rarely assigns extra labels when evidence for a second claim is weak.

3.1 Policy Misperception Themes

Reddit discussions of the BC policy revealed several recurring misconceptions. After excluding unclear or off-topic remarks, we identified four dominant themes. In 2023 and early 2024, two themes accounted for most identified misconceptions, including *'Policy-Target Misinterpretation'* and *'Impact*

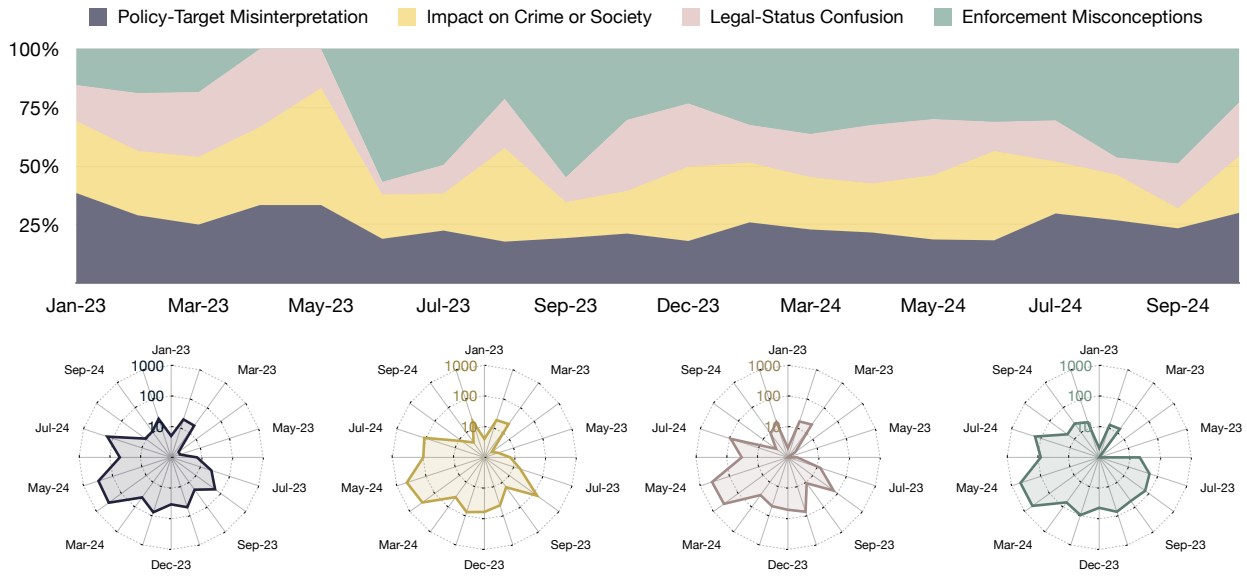

Figure 2: **Composition and dynamics of misperception themes over time (Jan 2023 to Sep 2024).** The top row is a percentage area chart. At the end of each month, the vertical axis represents the fraction of posts in each class, showing how their relative shares change over time. The bottom row presents radar plots on a logarithmic radial scale, which emphasize month-to-month intensity within each class and make intermittent peaks more visible. October 2024 is omitted because the collection window ends on 20 October 2024, which yields a partial month.

on Crime or Society'. The first refers to confusion about the scope of the exemption (e.g., some users assumed the policy legalizes drug selling or applies to drugs and contexts beyond what is actually allowed). The second theme involves claims that the policy causes more crime, public disorder, or other social harms and aligns with fears about its impact on communities.

Figure 2 illustrates the theme composition conditional on Labels 1-4, where the Unclear/No verifiable claim category is excluded from the denominator. In most months, the mix of discussions was fairly stable and was dominated by the two core themes above. However, a noticeable shift occurred in mid-2024. *'Legal-Status Confusion'* and *Enforcement Misconceptions* accounted for roughly 15% of all comments in June 2024. The increase was clearer in absolute volume and in reply engagement during the weeks after the May 2024 amendment. This uptick coincided with a significant policy amendment in May 2024, when Health Canada and the BC government narrowed the decriminalization pilot by prohibiting drug possession in all public spaces.[42] Since then, legal possession has been limited to private residences and certain designated sites only.[42] During that period, many Reddit users appeared unsure about what was actually allowed under 'decriminalization'. Some wondered whether the policy had been effectively reversed and others complained that 'police are still arresting people', which suggests widespread confusion about the policy's legal boundaries and how enforcement would proceed after the amendment.

By late 2024, the distribution of misconception themes began to revert to its earlier pattern, where the discussion of legality and enforcement issues became less prominent, and the conversation shifted back to focusing more on the policy's targets and its perceived societal impact. This suggests that the mid-2024 spike in legal/enforcement confusion was an episodic reaction to the policy change rather than a permanent refocusing of the discourse. In other words, public attention to legality and enforcement rose around the revision and fell in later months as the changes became clearer.

3.2 Bursts of Reply Engagement

To determine whether surges in enforcement-related misconceptions corresponded to engagement spikes, we identified periods when these comments received an unusually large number of replies. Figure 3 maps the weekly reply engagement premium for legality/enforcement misperception posts across subreddits and highlights four distinct 'burst' episodes in early to mid-2024. Each burst, outlined by a polygon in the figure, represents a cluster of weeks in which comments containing legal-status or enforcement misconceptions consistently received far more replies than other posts. All four engagement bursts occurred between February and July 2024. They ranged in duration from 2 to 5 weeks and involved simultaneous activity in multiple subreddits (between 2 and 5 communities in each burst), spanning both regional forums (e.g., `r/britishcolumbia`, `r/VictoriaBC`) and broader national forums (e.g., `r/canada`, `r/ontario`). Interestingly, we did not observe any comparable multi-week clusters of negative engagement, where misperception posts received significantly fewer replies than usual. This asymmetry suggests that when misconceptions about the policy arose, they tended to attract attention and spark debate rather than being ignored.

One prominent burst followed the early May 2024 policy revision. In the weeks immediately after the new public-use restrictions took effect, misinformed comments about the law's enforcement saw a surge of replies across several subreddits. Users were actively debating and clarifying the changes. For

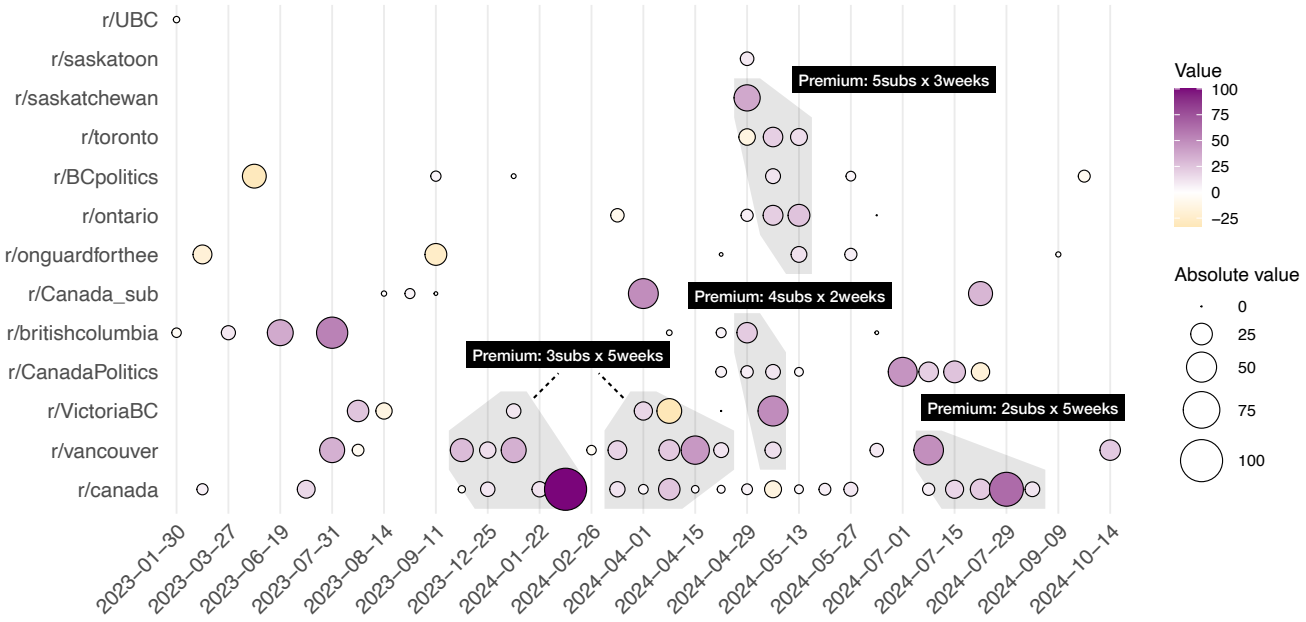

Figure 3: **Weekly reply premium by subreddit with multi-week × multi-subreddit bursts.** Points are subreddit-week observations, where point area encodes |premium| and color is centered at zero. Purple indicates a positive premium and orange a negative penalty. The polygons show contiguous bursts to highlight the patterns (e.g., "5 subreddits × 3 weeks").

example, some threads featured arguments about whether police could now ticket or arrest people using drugs in parks despite the decriminalization policy, while others had users expressing frustration or support regarding the reinstated penalties for public possession. This burst coincided with the larger share of legality and enforcement comments reported in RQ1. The May 2024 policy change shifted the discourse and increased its intensity, as many commenters expressed confusion and debated enforcement.

The other bursts in early 2024 similarly corresponded to moments of increased public interest or controversy. For instance, one cluster in February 2024 involved both local and national subreddits reacting to news stories about visible drug use in public spaces. This led to heated discussions on Reddit about whether decriminalization was 'to blame' and what should be done, with many misconceptions voiced about what police were or were not allowed to do. Another burst in spring 2024 followed public calls from some municipal leaders for stricter enforcement amid concerns about open drug use, which again generated Reddit threads full of claims and counterclaims about the policy's enforcement. In each case, comments that mischaracterized the legal or enforcement aspects of the policy became focal points for extended back-and-forth exchanges.

Across all detected bursts, comments containing legality or enforcement misconceptions enjoyed a clear engagement premium, where they consistently garnered more replies than other comments posted around the same time. Outside of these clustered periods, the reply premium for such misinformed comments was smaller and more sporadic. In other words, the intense engagement with legality/enforcement misconceptions was concentrated in specific windows when those issues were especially salient in public debate. From a policy perspective, identifying these windows is valuable because it shows when confusion about the law peaked and which online communities were most involved. This pattern creates an opportunity for targeted outreach or clarification.

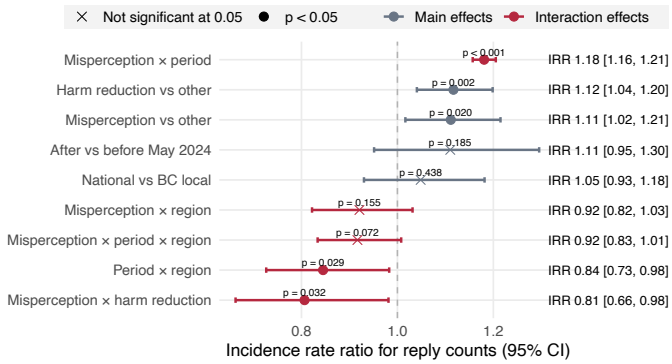

Figure 4: **Incidence rate ratios (IRR) for direct reply counts from a negative binomial regression.** Points show incidence rate ratios and bars show 95% confidence intervals. The model includes comment length, upvotes, hour, weekday, and subreddit fixed effects. Interaction terms link legal or enforcement misconceptions with policy period, region, and harm-reduction content

3.3 Harm reduction discourse and engagement with legal/enforcement misconceptions

Figure 4 summarizes adjusted associations between legal and enforcement misperceptions, harm-reduction content, policy period, region, and reply counts. Misperception comments received more replies than other comments after adjustment for comment length, upvotes, time-of-day, weekday, and subreddit

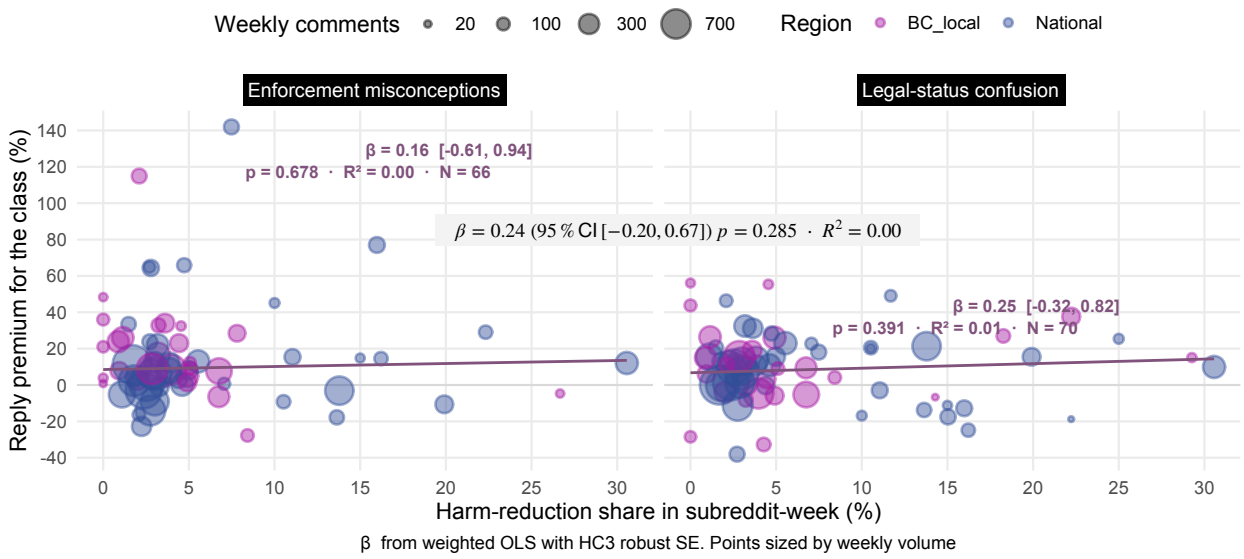

Figure 5: **Harm reduction discussion and engagement with policy misconceptions.** Each point represents a subreddit week and links the share of harm reduction content to the reply premium for legal status and enforcement misconceptions, and point size reflects weekly comment volume while color distinguishes BC local from national communities. The pattern shows that engagement with these misconceptions remains high across a wide range of harm reduction visibility during periods of policy concern.

fixed effects. The IRR for misperception versus other content was 1.11 (95% CI 1.02-1.21, $p = 0.020$), which corresponds to an 11% reply premium on average. Harm-reduction content also showed a reply premium, with an IRR of 1.12 (95% CI 1.04-1.20, $p = 0.002$). The period main effect was smaller and imprecise, with After vs before May 2024 at IRR 1.11 (95% CI 0.95-1.30, $p = 0.185$). In contrast, the misperception-by-period interaction was positive and large, with Misperception × period at IRR 1.18 (95% CI 1.16-1.21, $p < 0.001$). This interaction indicates a larger reply premium for misperception comments in the post-revision period.

Regional differences were small after adjustment. The misperception-by-region term was close to null, with IRR 0.92 (95% CI 0.82-1.03, $p = 0.155$). The period×region interaction was below one, with IRR 0.84 (95% CI 0.73-0.98, $p = 0.029$), which indicates that the post-revision change differed by region. The three-way misperception×period×region term had limited precision, with IRR 0.92 (95% CI 0.83-1.01, $p = 0.072$). Interactions in the error-bar plot highlight how these effects combine. The misperception×period term was positive, with misperception comments in the post-revision period drawing more replies than misperception comments in the pre-revision period, over and above the overall increase in replies. In contrast, the misperception×region interaction was close to null, suggesting that the average misperception reply premium was similar in BC-local and national subreddits once other variables were controlled. The period×region interaction showed smaller post-revision increases in reply counts in BC-local communities than in national ones. The misperception×period×region term had limited precision, and the overall pattern was similar across regions.

The interaction between misperceptions and harm reduction was negative. The IRR for the misperception×harm-reduction term was 0.81 (95% CI 0.66–0.98; $p = 0.029$). This implies that misperception comments and harm-reduction comments each received more replies on average. Misperception comments that also contained harm-reduction content showed a lower reply premium than misperception comments without harm reduction. The additional replies linked to misperceptions were smaller when the same posts referenced harm-reduction concepts. Model-based marginal means followed the same pattern. Predicted reply counts were lowest for comments with neither misperceptions nor harm-reduction content, higher for comments with only misperceptions or only harm-reduction content, and highest when misperceptions occurred in the post-revision period in national forums. However, within the subset of misperception comments, adding harm-reduction content did not substantially increase replies and slightly reduced the premium in some groups, consistent with the negative interaction term.

Figure 5 links, at the subreddit–week level, the share of harm-reduction content to the 'reply premium' for misperception comments. Across both classes combined, the volume-weighted ordinary least-squares slope was small and not statistically significant, where a 10% increase in the harm-reduction share was associated with an estimated 2.4 percentage-point increase in misperception reply premium (95% CI –2.0 to 6.7; $p = 0.285$; $R^2 \approx 0.00$). Class-specific panels showed similarly flat relationships. For Enforcement Misconceptions, the slope was 1.6 percentage points per 10 percentage-point increase in harm-reduction share (95% CI -6.1 to 9.4; $p = 0.678$). For *'Legal-Status Confusion'*, the slope was 2.5 percentage points (95% CI –3.2 to 8.2; $p = 0.391$). In both cases, weekly reply premiums varied widely across subreddit–weeks, but there was no clear tendency for weeks with more harm-reduction discussion to exhibit larger or smaller engagement gaps be-

tween misperception and other comments. These findings show that legal and enforcement misperceptions and harm-reduction posts attract slightly more replies than other comments at the individual-comment level after the May 2024 policy revision. However, week-to-week variation in community-level harm-reduction visibility does not consistently change the reply premium for misperceptions.

4 Discussion

4.1 Online Misconceptions as Implementation Signals

We interpret the online discourse around BC's decriminalization pilot as a set of implementation signals that can shape opioid-related harms. The comments on Reddit show how people explain visible drug use, how they understand police powers, and how they interpret the policy's purpose. Each interpretation can influence support for harm reduction services, willingness to seek care, and expectations about enforcement. The themes below connect these narratives to available evidence and outline communication implications for policy changes.

Perceived Disorder and Public Safety: Concern about public safety often relied on the visibility of drug use in shared spaces. Many commenters treated visible public use as direct evidence that decriminalization increased crime and overdoses, as illustrated by one comment: *"Since decriminalization was put in place, there has been more open drug use. More social disorder and crime. I get all of those things like mental health and housing availability and people's income but decriminalization made things worse."* Our findings reflect broader national trends in which decriminalization was also associated with perceptions of increased crime, social disorder, and overdoses. For instance, *the Canadians' Knowledge and Attitudes Around Drug Decriminalization* survey reported that 51% of respondents in 2023 and 53% in 2024 believed decriminalization would lead to greater harms, such as overdoses.[21] Similarly, 43% in 2023 and 48% in 2024 agreed that decriminalizing drugs would make their communities less safe.[21]

In contrast to these perceptions, enforcement data indicate that decriminalization achieved immediate enforcement goals. During the initial exemption period in BC, possession offences and sub-threshold seizures declined sharply relative to pre-policy baselines, consistent with reduced criminalization of personal possession.[43,44,45,46,47] This pattern is consistent with reduced criminal enforcement for personal possession, which is a core pathway through which decriminalization is expected to reduce criminal justice harms. Evidence from other jurisdictions also cautions against interpreting concurrent overdose increases as a direct policy effect. In Oregon and Washington, quasi-experimental analyses found no clear increase in overdose deaths over the first year of decriminalization, while arrests for possession declined substantially.[48,49] In Oregon, models that accounted for the rapid spread of fentanyl no longer attributed overdose mortality to decriminalization.[50] Qualitative work in the same period describes fentanyl's growing dominance in local markets, which offers a plausible explanation for rising fatalities that is independent of legal changes.[51] Decriminalization may therefore affect visibility and public expectations, while overdose trajectories track toxic supply dynamics and structural vulnerability more than legal status.[46,48,50]

Misunderstandings About Enforcement and Consequences: A common theme was the belief that decriminalization eliminated consequences for possession and ended police intervention, as one argued: *"The concept that someone can be fined for having a drink in public but you can do drugs... with no consequences makes no sense."* National survey results suggest that this misunderstanding is widespread. In 2023, 60% of respondents and, in 2024, 65% believed it is now legal to possess any drug and that police no longer stop people carrying illegal substances.[21] Studies of people who use drugs describe similar confusion about what decriminalization changes in practice. Participants often conflated decriminalization with legalization and reported uncertainty about when police still intervene.[26] This uncertainty can create two forms of risk. One risk is a false sense of immunity to enforcement. Another risk is continued fear and mistrust that discourages engagement with services. The most recent *Quarterly Decriminalization Data* reports indicate ongoing police intervention, with more than half of possession offences between May 2024 and January 2025 occurring in public spaces such as streets, roads, highways, and parking lots.[52] Public communication that distinguishes decriminalization from legalization and explains remaining enforcement pathways is therefore part of implementation, not an optional add-on.

Confusion About Permissible Locations: Another cluster of comments focused on where use or possession would be permitted. Many commenters assumed that decriminalization applied in all public spaces and raised concerns about child-focused environments, as one stated: *"This was so stupid. Why would you ever make that legal in public places? There's a time and place for everything. Smoking meth at the bus stop in front of kids is not it."* The legal framework does not match this interpretation. The original exemption in January 2023 already prohibited possession in certain spaces, and amendments in September 2023 expanded restrictions in child-focused areas.[52] A revised exemption in May 2024 narrowed permitted settings further and limited possession to private residences, legal sheltering sites, and supervised consumption services.[52] Frequent changes to the exemption likely increased confusion for the public and for people most exposed to enforcement, which increases opportunities for misinformation to circulate. Effective communication for this type of misunderstanding requires place-based clarity. Simple descriptions of permitted and non-permitted settings can reduce uncertainty. Reduced uncertainty can also lower avoidable contact with police that is driven by misunderstanding rather than trafficking or violence.

Moral Framing and Policy Purpose: A separate theme involved moral interpretations of the policy's intent. Some commenters framed decriminalization as government endorsement of drug use, rather than a measure aimed at reducing stigma and linking people with care, as one noted: *"The government encouraging drug use has been extremely triggering as someone trying to stay away from it."* Media framing provides context

for this response. A recent scoping review reports that media portrayals often characterize people who use drugs as *criminal*, *violent*, or *dangerous*, which can shift policy preferences toward punitive approaches rather than treatment.[53] Opioid-related coverage also often centers on scenes of public overdoses and the presence of police or first responders, which can frame the crisis as one warranting a criminal justice response.[54]

Within this environment, decriminalization can be interpreted as permissiveness unless communication links the policy to clear public health goals and service pathways. Reported increases in the proportion of people who use drugs accessing overdose prevention and supervised consumption services provide a concrete metric for public communication.[52] Decriminalization also requires treatment and harm reduction capacity that is visible and accessible. Portugal's model is often cited because it paired the end of criminal penalties with investments in care. Portuguese service entry also begins with assessment of socioeconomic needs such as housing and family support, and harm reduction services provide supplies alongside information on treatment options.[55]

Engagement Bursts and Communication Timing: Online engagement with legal-status and enforcement misconceptions intensified around key moments in the policy timeline. The May 2024 amendment, which reintroduced penalties for drug possession in public spaces, concentrated confusion and debate. Comments that misrepresented the legal scope of the exemption also attracted larger reply chains, which indicates that misinformation can become a focal point for public argument.

Administrative records after the May 2024 revision show that in the subsequent eight months, roughly half of drug possession tickets in BC were issued in public locations where possession had become re-criminalized.[27,42] This pattern suggests that many people did not fully understand the new rules at the point when enforcement risk increased. In fast-changing policy settings, timely clarification is an implementation component that can reduce avoidable contact with police and support access to health services. Detection of misconception-driven spikes provides a practical targeting tool for public health communication. When online discussion concentrates confusion in identifiable weeks and community spaces, agencies and community organizations can deploy short, consistent messages in the same venues. After major revisions, such messaging can clarify where possession is decriminalized, what police can still do, and where people can access harm reduction supports. Online discussion is where decriminalization's rules get interpreted in real time. When misconceptions about scope and enforcement dominate, they can weaken support for evidence-based services and skew how the reform is judged.

4.2 Harm Reduction Legitimacy in Online Debate

To reduce opioid-related harms, harm reduction policy requires evidence-based design and public legitimacy. In online spaces, that legitimacy is negotiated through narratives about safety, responsibility, and policy intent. This negotiation has practical stakes because naloxone distribution and supervised consumption services have strong evidence of benefit, and access to these interventions depends on sustained political support.[56,57,58]

Public opinion data indicate a broad base of support for harm reduction in Canada. A national survey found that about 64% of Canadians support harm reduction approaches.[59] Earlier research in Canada, the UK, and Australia also reported majorities backing needle programs and supervised consumption.[60] This support does not always translate into stable policy or durable institutional commitment. In one systematic review, public health experts noted that survey findings of strong support were 'ignored by some policy-makers and media', alongside negative portrayals of safer drug use practices.[60] Consistent with this pattern, policy analyses have described Canadian harm reduction frameworks as symbolic and rhetorical rather than substantive.[59] Work on harm reduction debates also indicates that evidence uptake is often filtered through moral and emotional commitments, which can limit the influence of effectiveness data alone.[61] Media effects research offers a mechanism for how selective coverage can shape public attitudes and policy preferences over time. Cultivation processes link sustained exposure to recurring narratives with changes in social reality beliefs, particularly when individuals have limited direct experience with an issue.[62] Recent analyses of Canadian news similarly found that coverage of youth substance use often emphasized generalized 'harms on the rise' and individual-level solutions, with limited attention to social context.[63] In this environment, visible disorder narratives can become a dominant policy lens, even when the evidence base for harm reduction outcomes is substantial.[57,56]

Online discourse adds another layer of complexity. Social media and forums can enable peer support and rapid circulation of harm reduction practices, but they can also spread misleading claims and stigma.[64] Health communication scholarship describes this dynamic as part of an 'infodemic', where misleading information can contribute to confusion, risk-taking behaviours, and reduced trust in health authorities.[65,66] The language used in these discussions also matters for policy preferences. In a randomized study with clinicians, exposure to the label 'substance abuser' increased endorsement of punitive responses compared with describing the person as having a 'substance use disorder'.[67] This finding supports the concern that blame-focused discourse can shift conversations away from care, even in settings intended to be supportive.

These patterns indicate a concrete communication opportunity inside harm reduction conversations. Stigma reduction strategies that reduce social distance, including lived-experience storytelling and non-stigmatizing language, can support more durable acceptance of harm reduction policy. Misinformation research also supports upstream prevention. Inoculation and prebunking approaches that teach common manipulation tactics can reduce susceptibility to misinformation, and they can scale through short, digital formats.[68,69] For opioid policy, this supports a two-part strategy. First, prebunking and clear policy explainers can establish baseline understanding before high-confusion moments. Second, targeted debunking can respond to high-velocity false claims when they surge,

consistent with infodemic management guidance.[65,66]

4.3 Operationalizing Rapid Harm Reduction and Early Warning on Social Platforms

We studied reply behaviour to identify when misconceptions concentrate attention in opioid policy debates. In our data, comments that referenced harm reduction attracted more replies than other comments. Comments that misrepresented the legal status or enforcement of decriminalization attracted more replies, with the largest increases after the May 2024 exemption revision. This pattern indicates a predictable implementation window for real-time clarification and service signposting.[52,27,42] During that window, clear place-based rules and service signposting can reduce misunderstanding and avoidable enforcement exposure.[52,27,42]

Two findings help interpret what this engagement means for harm reduction implementation directly. Harm reduction language predicted higher reply counts, which indicates that peers use these threads for practical support. The interaction between misperception and harm reduction content was negative, which suggests that threads can shift toward clarification when guidance appears. Prior forum research describes similar norms, where users share dosing cautions, safer use practices, and mutual aid in disruptive periods.[39,40] Our burst detection results extend this interpretation from individual threads to community-level signals. Clusters of elevated reply premiums appeared across multiple subreddits in early to mid-2024, with a clear spike after the May 2024 revision. Community-level harm reduction visibility did not predict these reply premiums, which indicates that confusion can spike even when harm reduction talk is present. Prior surveillance research supports this early signal, where opioid discussions on social platforms provide lead time for overdose trends and emerging substance exposures.[70,71] Policy relevance is largest when legal context and overdose risk diverge in the same period. Evaluations in Oregon and Washington show large declines in possession arrests after decriminalization, while overdose trends track fentanyl market dynamics more than enforcement levels.[48,49,50,51] Portugal also illustrates that decriminalization benefits depend on accessible treatment and harm reduction capacity, which digital signposting can help coordinate.[55]

To translate these dynamics into fewer harms, agencies can treat platform discourse as a routine implementation surface, guided by ethical social listening standards.[72] Communication should align with exemption revisions and emphasize location rules and remaining enforcement powers.[52,27,42] Prebunking that distinguishes decriminalization from legalization can prime readers before high-attention bursts online.[68,69] Partnerships with moderators can keep naloxone, opioid agonist treatment, supervised consumption, and safer supply resources visible when new threads form.[39,40] A final component connects online signals to offline logistics when agencies review indicators and adjust response capacity.[70,71] Such a workflow positions social platforms as a scalable harm reduction interface that complements public support for evidence-based responses.[59,60,64,63]

4.4 Limitations

Several limitations should be considered when interpreting the online signals captured in this study and when generalizing beyond Reddit. Our corpus includes 22,131 comments posted between January 2023 and October 2024, and activity is concentrated in large subreddits. This study used a keyword search for the phrase 'BC Decriminalization', which means relevant threads that did not include this phrase, or posts in other languages, may have been missed. As a result, our analysis describes conversations within selected online communities, rather than population views in British Columbia or Canada. Moreover, omitting user identifiers protected privacy, but it prevented assessment of bots, repeated accounts, or geographic location.

To maintain labeling reliability, we limited the misconception taxonomy to four checkable domains and an 'unclear' category for uncertain cases. Because many comments contained no checkable claim, trend estimates exclude stigma, moral framing, and uncertainty when they were expressed without falsifiable statements. Zero-shot classification with general-purpose models can misread sarcasm, idioms, or context that spans multiple comments. To mitigate this, we conducted manual review of disagreement cases to select a primary model. However, this process does not estimate corpus-wide error rates. Some labels require judgment about evidentiary strength, which can vary when commenters cite news reports or personal experience.

Our proposed reply-premium metric identifies weeks when misconception comments attract attention, but it cannot distinguish between correction and amplification. Engagement was measured using direct reply counts, so it does not capture deeper thread dynamics or user-to-user diffusion beyond first-order responses. Temporal bursts may also reflect external events, media coverage, or moderation practices, none of which were modeled as time-varying covariates in our study. Moreover, we used the policy revision as a reference point for comparison, but causal attribution is limited in observational social data. Future work can extend the collection window, incorporate additional platforms, and link online discourse to administrative and health indicators on finer time scales.

5 Conclusion

We analyzed 22,131 Reddit comments to quantify how policy misunderstanding concentrates attention during decriminalization implementation in British Columbia. Zero-shot multi-label classification mapped each comment to a fixed misconception taxonomy and identified policy-purpose and social-harm misconceptions as dominant. Confusion about legality and enforcement increased after May 2024, and a post-revision reply premium made those threads more contested. After covariate adjustment, comments with legal or enforcement misconceptions received about 11% more replies, and the premium grew after May 2024. Weekly reply premiums clustered into four multi-subreddit bursts from February to July 2024 and concentrated disputes in a few venues. Within misconception

threads, harm reduction references carried an engagement premium and reduced reply spikes, which supports plain explanations and service links. Because policy revisions predict confusion peaks, communications should state place-based possession rules and clarify ongoing police authority. Burst detection can guide when and where to post brief messages that link audiences to naloxone, supervised consumption, and opioid agonist treatment. Open code and stable label definitions make the pipeline portable for teams that need measurable feedback during policy change without custom training data. This design treats debate as policy telemetry and supports rapid and low-burden communication planning for implementers.

Abreviations

BC	British Columbia
AB	Alberta
ON	Ontario
LLM	Large Language Model
IRR	Incidence Rate Ratios
CI	Confidence Interval
SD	Standard Deviation
IQR	Interquartile range

Author's contributions

SH, MS, CPY, KS, and MN contributed to data collection and the conceptualization of the study. SH contributed to the development of the LLM pipeline. MS contributed to the implementation and validation of the data labelling process and wrote the initial draft. KS, CD, and MN reviewed, edited, and provided feedback on the manuscript. ZS conceptualized and designed the study, conducted the formal data analysis and visualization, and wrote the final version of the manuscript. ZS was responsible for funding acquisition and overall supervision of the project.

Funding

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Canada Research Chairs Program and Discovery Grant (RGPIN-2025-07037).

Data availability

The data used in this study, including the raw dataset and labeled datasets produced by each large language model used in the methods, are publicly available in the provided GitHub repository.[25]

Declarations

Ethics approval and consent to participate

This study analyzed publicly available Reddit comments and did not involve direct interaction with human participants. Because this research used open data and included no identifiers, ethics approval and participant consent were not required.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

[1] Public Health Agency of Canada. Opioid- and stimulant-related harms in canada: Key findings (2024).

[2] Ballantyne, J. C. & Mao, J. Opioid therapy for chronic pain. *New England Journal of Medicine* **349**, 1943–1953 (2003).

[3] Volkow, N. D. & McLellan, A. T. Opioid abuse in chronic pain—misconceptions and mitigation strategies. *New England Journal of Medicine* **374**, 1253–1263 (2016).

[4] Compton, W. M. & Volkow, N. D. Major increases in opioid analgesic abuse in the united states: concerns and strategies. *Drug and alcohol dependence* **81**, 103–107 (2006).

[5] White, J. M. & Irvine, R. J. Mechanisms of fatal opioid overdose. *Addiction* **94**, 961–972 (1999).

[6] Rudd, R. A. Increases in drug and opioid-involved overdose deaths—united states, 2010-2015. *MMWR. Morbidity and mortality weekly report* **65** (2016).

[7] Canadian Substance Use Costs and Harms Scientific Working Group. Canadian substance use costs and harms 2007–2020. Tech. Rep., Canadian Centre on Substance Use and Addiction & Canadian Institute for Substance Use Research (2023).

[8] Government of Canada. The canadian drugs and substances strategy (2024).

[9] Canadian Substance Use Costs and Harms Scientific Working Group). Canadian substance use costs and harms (csuch) 2007-2020. Report, Canadian Centre on Substance Use and Addiction, Ottawa, ON (2023).

[10] Health Canada. Stigma around drug use (2024). URL https://www.canada.ca/en/health-canada/services/opioids/stigma.html.

[11] Voon, P. *et al.* Pain as a risk factor for substance use: a qualitative study of people who use drugs in british columbia, canada. *Harm reduction journal* **15**, 1–9 (2018).

[12] Scheim, A. I. *et al.* Impact evaluations of drug decriminalisation and legal regulation on drug use, health and social harms: a systematic review. *BMJ Open* **10**, e035148 (2020). URL http://dx.doi.org/10.1136/bmjopen-2019-035148.

[13] Ontario Agency for Health Protection and Promotion (Public Health Ontario). Scan of evidence and jurisdictional approaches to the decriminalization of drugs. Report, Public Health Ontario, Toronto, ON (2022).

[14] Oregon State Legislature. House Bill 4002: An Act Relating to the addiction crisis in this state; and declaring an emergency (2024).

[15] Government of British Columbia. Decriminalizing people who use drugs in b.c. https://www2.gov.bc.ca/gov/content/overdose/decriminalization (2025).

[16] BC Centre for Disease Control. Decriminalization in B.C. (2023). URL https://www.bccdc.ca/health-info/prevention-public-health/decriminalization-in-bc.

[17] Russell, C. *et al.* Exploring the early impacts of drug decriminalization on harm reduction and opioid agonist treatment service operations and delivery in british columbia: insights from key informant interviews. *BMC Public Health* **26**, 157 (2026).

[18] Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).

[19] Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).

[20] Jesseman, R. & Payer, D. Decriminalization: Options and evidence. Tech. Rep., Canadian Centre on Substance Use and Addiction (CCSA), Ottawa, ON (2018). URL https://www.ccsa.ca/sites/default/files/2019-04/CCSA-Decriminalization-Controlled-Substances-Policy-Brief-2018-en.pdf.

[21] Health Canada. Canadians' knowledge and attitudes around drug decriminalization: Results from a public opinion research survey (2023).

[22] Shakeri Hossein Abad, Z. *et al.* Digital public health surveillance: a systematic scoping review. *NPJ digital medicine* **4**, 41 (2021).

[23] Tourangeau, R. & Yan, T. Sensitive questions in surveys. *Psychological Bulletin* **133**, 859–883 (2007).

[24] Krumpal, I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* **47**, 2025–2047 (2013).

[25] HIVE Lab. Opioid_Decriminalization: Code and dataset for detecting misconception surges in opioid policy. https://github.com/HIVE-UofT/Opioid_Decriminalization (2026).

[26] Greer, A., Xavier, J., Loewen, O. K., Kinniburgh, B. & Crabtree, A. Awareness and knowledge of drug decriminalization among people who use drugs in british columbia: a multi-method pre-implementation study. *BMC Public Health* **24**, 407 (2024).

[27] Government of British Columbia, Ministry of Mental Health & Addictions. Decriminalization-data report to health canada, february 2023 to january 2025. Tech. Rep. Third quarterly report, Government of British Columbia (2025).

[28] Dineen, K. K. Definitions matter: a taxonomy of inappropriate prescribing to shape effective opioid policy and reduce patient harm. *U. Kan. L. Rev.* **67**, 961 (2018).

[29] Gilardi, F., Alizadeh, M. & Kubli, M. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**, e2305016120 (2023).

[30] Kazari, K., Chen, Y. & Shakeri, Z. Scaling public health text annotation: Zero-shot learning vs. crowdsourcing for improved efficiency and labeling accuracy. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4 (IEEE, 2025).

[31] Guo, Y., Ovadje, A., Al-Garadi, M. A. & Sarker, A. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association* **31**, 2181–2189 (2024).

[32] Fachada, N., Fernandes, D., Fernandes, C. M., Ferreira-Saraiva, B. D. & Matos-Carvalho, J. P. Gpt-4.1 sets the standard in automated experiment design using novel python libraries (2025). 2508.00033.

[33] Comanici, G. *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025). 2507.06261.

[34] Zhang, M.-L. & Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* **26**, 1819–1837 (2014).

[35] Tsoumakas, G. & Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* **3**, 1–13 (2007).

[36] Touvron, H. *et al.* Llama: Open and efficient foundation language models (2023). 2302.13971.

[37] Grattafiori, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[38] Kleinberg, J. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery* **7**, 373–397 (2003).

[39] Soussan, C. & Kjellgren, A. Harm reduction and knowledge exchange: a qualitative analysis of drug-related internet discussion forums. *Harm Reduction Journal* **11**, 25 (2014).

[40] Bunting, A. M. *et al.* Socially-supportive norms and mutual aid of people who use opioids: An analysis of reddit during the initial COVID-19 pandemic. *Drug and Alcohol Dependence* **222**, 108672 (2021).

[41] Cameron, A. C. & Miller, D. L. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* **50**, 317–372 (2015).

[42] Health Canada. Personal possession of small amounts of certain illegal drugs in british columbia (2024). Health Canada news release, May 2024.

[43] Government of British Columbia, M. o. M. H. & Addictions. Decriminalization: Data report to health canada, february 2023-january 2024. Tech. Rep., Government of British Columbia, Ministry of Mental Health and Addictions (2024). URL https://www2.gov.bc.ca/assets/gov/overdose-awareness/data_report_to_health_canada_may_2024.pdf. Accessed 2025-08-18.

[44] Government of British Columbia, M. o. M. H. & Addictions. Decriminalization: Data report to health canada, february 2023-april 2024. Tech. Rep., Government of British Columbia, Ministry of Mental Health and Addictions (2024). URL https://www2.gov.bc.ca/assets/gov/overdose-awareness/data_report_to_health_canada_august_2024.pdf. Accessed 2025-08-18.

[45] Government of British Columbia, M. o. M. H. & Addictions. Decriminalization: Data report to health canada, february 2023-july 2024. Tech. Rep., Government of British Columbia, Ministry of Mental Health and Addictions (2024). URL https://www2.gov.bc.ca/assets/gov/overdose-awareness/data_report_to_health_canada_november_2024.pdf. Accessed 2025-08-18.

[46] Government of British Columbia, M. o. M. H. & Addictions. Decriminalization: Data report to health canada, february 2023-october 2024. Tech. Rep., Government of British Columbia, Ministry of Mental Health and Addictions (2025). URL https://www2.gov.bc.ca/assets/gov/overdose-awareness/data_report_to_health_canada_february_2025.pdf. Accessed 2025-08-18.

[47] Vancouver Police Department. Drug possession seizures decline during decriminalization pilot. https://vpd.ca/news/2024/03/26/vpd-drug-seizures-decline-during-decrim-pilot/ (2024). Accessed 2025-08-18.

[48] Joshi, S. *et al.* One-year association of drug possession law change with fatal drug overdose in oregon and washington. *JAMA Psychiatry* **80**, 1277–1283 (2023).

[49] Davis, C. S., Joshi, S., Rivera, B. D. & Cerdá, M. Changes in arrests following decriminalization of low-level drug possession in oregon and washington. *International Journal of Drug Policy* **119**, 104155 (2023).

[50] Zoorob, M. J., Park, J. N., Kral, A. H., Lambdin, B. H. & del Pozo, B. Drug decriminalization, fentanyl, and fatal overdoses in oregon. *JAMA Network Open* **7**, e2431612 (2024).

[51] Shin, S. S. *et al.* "it wasn't here, and now it is. it's everywhere": fentanyl's rising presence in oregon's drug supply. *Harm Reduction Journal* **19**, 76 (2022).

[52] Government of British Columbia. Decriminalization: Data report to Health Canada (february 2023-january 2025). Tech. Rep., Government of British Columbia (2025).

[53] Bosworth, K. T. *et al.* Analysing media portrayals of people with substance use disorder and addiction: A scoping review. *Cultures of Science* **7**, 126–141 (2024).

[54] Knaak, S., Mercer, S., Christie, R. & Stuart, H. Stigma and the opioid crisis. *Mental Health Commission of Canada. Retrieved from https://www. mentalhealthcommission. ca/sites/default/files/2019-07/Opioid_Report_july_2019_eng. pdf* (2019).

[55] Hughes, C. E. & Stevens, A. What can we learn from the portuguese decriminalization of illicit drugs? *The British Journal of Criminology* **50**, 999–1022 (2010).

[56] McDonald, R. & Strang, J. Are take-home naloxone programmes effective? systematic review utilizing application of the bradford hill criteria. *Addiction* **111**, 1177–1187 (2016).

[57] Potier, C., Laprévote, V., Dubois-Arber, F., Cottencin, O. & Rolland, B. Supervised injection services: what has been demonstrated? a systematic literature review. *Drug and Alcohol Dependence* **145**, 48–68 (2014).

[58] Marshall, B. D. L., Milloy, M.-J., Wood, E., Montaner, J. S. G. & Kerr, T. Reduction in overdose mortality after the opening of north america's first medically supervised safer injecting facility: a retrospective population-based study. *The Lancet* **377**, 1429–1437 (2011).

[59] Wild, T. C. *et al.* Public support for harm reduction: A population survey of canadian adults. *PloS one* **16**, e0251860 (2021).

[60] Tzemis, D., Campbell, J., Kuo, M. & Buxton, J. A. A cross-sectional study of public attitudes towards safer drug use practices in british columbia, canada. *Substance Abuse Treatment, Prevention, and Policy* **8**, 40 (2013).

[61] Zampini, G. F. Evidence and morality in harm-reduction debates: can we use value-neutral arguments to achieve value-driven goals? *Palgrave Communications* **4**, 62 (2018).

[62] Ramondt, S. & Ramírez, A. S. Fatalism and exposure to health information from the media: examining the evidence for causal influence. *Annals of the International Communication Association* **41**, 298–320 (2017).

[63] Goodyear, T. *et al.* Taking stock of youth substance use portrayals: A critical content analysis of canadian news media, 2016-2024. *Social Science & Medicine* 118188 (2025).

[64] Mittal, S. *et al.* Exposure to content written by large language models can reduce stigma around opioid use disorder. *npj Artificial Intelligence* **1**, 46 (2025).

[65] World Health Organization. Infodemic. Web page (n.d.). URL https://www.who.int/health-topics/infodemic. Accessed 2026-01-10.

[66] Borges do Nascimento, I. J. *et al.* Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization* **100**, 544–561 (2022).

[67] Kelly, J. F. & Westerhoff, C. M. Does it matter how we refer to individuals with substance-related conditions? a randomized study of two commonly used terms. *International Journal of Drug Policy* **21**, 202–207 (2010).

[68] Lewandowsky, S. & van der Linden, S. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology* **32**, 348–384 (2021).

[69] Roozenbeek, J., van der Linden, S. & Nygren, T. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review* **1** (2020).

[70] Smith, D. A. *et al.* Monitoring the opioid epidemic via social media discussions. *npj Digital Medicine* **8**, 284 (2025).

[71] Barenholtz, E. *et al.* Online surveillance of novel psychoactive substances (nps): Monitoring reddit discussions as a predictor of increased nps-related exposures. *International Journal of Drug Policy* **98**, 103393 (2021).

[72] World Health Organization. Social listening in infodemic management for public health emergencies: guidance on ethical considerations. Tech. Rep., World Health Organization, Geneva (2025). URL https://www.who.int/publications/i/item/9789240108202. Accessed 2026-02-1.