SUPPLEMENTARY MATERIALS

Supplementary Analyses and Materials for Emotions, Context, and Substance Use in Adolescents: A Large Language Model Analysis of Reddit Posts

Table S1. Keyword list used for substance-related filtering.

Substance Category	Keywords
Alcohol	alcohol, drunk
Tobacco/Nicotine	tobacco, cigarettes, JUUL, vape, pens, nicotine
Cannabis	marijuana, weed, THC oils, dabs, leaf
Sedatives	sedatives, xanax, Ambien
Stimulants	stimulants, Adderall, cocaine
Hallucinogens	hallucinogens, LSD, PCP
Opioids	opioids, heroin, Vicodin, Percocet

* Note: Keywords were matched using case-insensitive regular expressions with word boundaries to reduce false positives. Several keywords (e.g., “weed,” “pen,” “leaf”) may have non-substance-related meanings in general discourse. To address this ambiguity, keyword matching was applied conservatively and followed by manual review, in which only posts with clear semantic references to substance use were retained. The validity of the keyword filtering step was further evaluated through manual annotation, as reported in the Methods.

Table S2. Examples of ambiguous or non-substance keyword matches identified during manual review.

Ambiguity Type	Keyword(s)	Description of Non-Substance Usage	Human Label (Substance_Use)	Handling
Lyrics/quoted text	alcohol, weed	Keywords appeared within song lyrics or quoted creative text without indicating personal substance use	0	Excluded from positive substance-use cases
Metaphorical language	weed	Keyword used metaphorically (e.g., “weed out”) rather than referring to cannabis	0	Excluded during manual review
Object reference	pen	Keyword referred to a physical object (e.g., writing instrument) rather than a vaping device	0	Excluded during manual review
Public health/pandemic context	alcohol	Keyword referred to alcohol used for sanitization or COVID-related discussion rather than consumption	0	Excluded during manual review

Table S3. Human validation of LLM-generated contextual annotations.

Context Dimension	Correct (Human = 1)	Total Evaluated	Validation Accuracy	F1 Score	Cohen’s κ
Family influence	286	300	0.953	0.901	0.871
Peer influence	287	300	0.957	0.971	0.889
School environment	293	300	0.977	0.885	0.872

Table S4. Manual annotation schema and label definitions.

Label Name	Values	Description
human_substance_use	{0, 1}	Whether the post explicitly expresses substance use (1 = explicit mention; 0 = no explicit substance use).
human_emotion_context	{0, 1}	Whether the LLM-assigned primary emotion accurately reflects the emotional content of the post.
human_context_family	{0, 1}	Whether the LLM’s family influence annotation is correct (1 = correct, 0 = incorrect), regardless of whether family influence is present or absent.
human_context_peer	{0, 1}	Whether the LLM’s peer influence annotation is correct (1 = correct, 0 = incorrect).
human_context_school	{0, 1}	Whether the LLM’s school environment annotation is correct (1 = correct, 0 = incorrect).

Note: Manual labels reflect human judgments of annotation correctness rather than independent re-annotation of ground-truth categories.

Table S5. Summary of manual annotation outcomes (n = 300).

Annotation Type	Dimension	Correct (1)	Total	Proportion Correct
Keyword filtering	Substance use	270	300	0.900
Context	Family influence	286	300	0.953
Context	Peer influence	287	300	0.957
Context	School environment	293	300	0.977
Emotion	Primary emotion	273	300	0.910

Note: Values indicate the proportion of LLM-generated annotations judged as correct by human reviewers.

Table S6. Illustration of manual validation fields (schematic examples).

LLM Output Field	LLM Annotation Type	Human Label Field	Human Judgment Meaning
emotion = fear	LLM emotion label	human_emotion_context = 1	LLM emotion correctly reflects the post's emotional content
emotion = anger	LLM emotion label	human_emotion_context = 0	LLM emotion does not accurately reflect the post's emotional content
Family Influence = none	LLM context (family)	human_context_family = 1	Correct absence of family influence
Family Influence = [1] family influence	LLM context (family)	human_context_family = 1	Correct identification of family influence
Peer Influence = [1] peer influence	LLM context (peer)	human_context_peer = 0	Incorrect identification of peer influence
School Environment = none	LLM context (school)	human_context_school = 0	Missed school-related context

Note: Examples are schematic representations illustrating how human-coded correctness labels correspond to LLM-generated annotations. No verbatim post text or model-generated rationale is shown to reduce re-identification risk.

Box S1. Example prompt for emotion annotation.

System prompt
You are an emotionally intelligent and empathetic agent. You will be given a piece of text, and your task is to identify the emotion expressed by the writer of the text. You are only allowed to make one selection from the following emotions, and do not use any other words: joy, guilt, anger, disgust, fear, sadness, shame.
User input
Document:
[Post content]
Output constraint
One label selected from: *joy, guilt, anger, disgust, fear, sadness, shame*.
If no clear emotion is expressed, the output is assigned *neutral*.

Note. Example prompt used for LLM-based emotion annotation following the ISEAR emotion framework. The placeholder *[Post content]* indicates the concatenated title and selftext of each post. The model was instructed to return a single primary emotion or *neutral* if no specific emotion was clearly expressed.

Box S2. Example prompts for contextual annotation.

Family influence prompt
You will receive a document related to adolescent behaviors. Identify any relevant topics that relate to family influence on substance use, such as parenting style, parental attitudes towards substances, family conflicts, or parental substance use patterns. If the document reflects family-related themes, output "[1] Family Influence" and briefly describe the relevant family-related context in the document. Otherwise, if no family influence is observed, return "None."
Document:
[Post Content]
Peer influence prompt
You will receive a document related to adolescent behaviors and substance use. Identify any relevant topics that indicate peer influence, such as peer pressure, social exclusion, group behavior at gatherings, or friends' substance use behaviors. If peer influence themes are present, output "[1] Peer Influence" with a brief description of the relevant peer-related context. If not, return "None."
Document:
[Post content]
School environment prompt
You will receive a document related to adolescent behaviors and substance use. Identify any relevant topics associated with the school environment, including school culture, teacher-student relationships, academic pressure, and student satisfaction with school life. If school-related factors are mentioned, output "[1] School Environment" with a brief description. If not, return "None."
Document:
[Post content]

Note. Example prompts used for LLM-based contextual annotation. Each contextual dimension (family, peer, and school) was evaluated independently using a dedicated prompt. The placeholder *[Post content]* indicates the concatenated title and selftext of each post.