<Project 1. Safety labelling & Policy interpretation>

[1] Executive Summary

This project is designed to simulate safety content reviewing and labeling. By defining safety policy categories, building decision rules, and applying them to 50 sample contents generated by GenAI, I am demonstrating my ability to interpret complex guidelines, make consistent labeling decisions, and document rationales in a structured way.

[2] Methodology

1. Safety categories: Developed clear definitions and examples of 9 categories (Violence, Hate Speech, Adult Nudity, Dangerous Acts, Minor Safety, Suicide & Self-harm, Bullying & Harassment, Illegal Activities & Regulated goods, and Fraud & Scam).
2. Decision trees: Constructed 2 logic flows for Minor Safety and Hate Speech which carry the highest operational and reputational risk and therefore benefit from structured, consistent labeling decisions.
3. Sample generation: Created 50 sample cases of diverse text and video descriptions that have both clear violations and ambiguous edge cases.
4. Policy application: Assigned category, severity level, action, and the rationale of justifying how policy was applied.
5. Patterns & insights: Examined category frequency, risk trends, and ambiguity challenges for reporting.

[3] Safety Categories

1) Violence – contents that depict, threaten, or encourage physical harms, violence, or injuries to people or animals. ("I'll beat you up", physical fights, bloods, animal abuse, attacks…etc.)
*Allowed: clearly fictional in movies/games, educational context, emergency response.

2) Hate Speech – contents that attacks, discriminates, or incites hatred against a protected group such as in race, ethnicity, nationality, religion, gender, sexual orientation, disability. ("I hate Japanese.")
*Allowed: insults toward non-protected groups such as games, political/social issues, an individual.

3) Adult Nudity & Sexual Content – contents that depict sexual acts, explicit nudity, sexual arousal, or sexualized positioning. Any sexual content involving minors are strictly prohibited. (pornographic materials, intentional focus on sexual body parts, promotion of adult content platforms, featuring minors)
*Allowed: normal swimwear, medical/educational contents, romantic/affectionate behavior without sexual intent.

4) Dangerous Acts – contents that display, encourage, or instruct high risk or harmful activities without safety precautions. (climbing/hanging from rooftops, balconies, or moving vehicles, handling dangerous items, risky challenges, weapon demonstrations, harmful encouragement)
*Allowed: performed by professionals with safety gear, sports with regulated rules, educational content

5) Minor Safety – contents involving anyone under 18 that include risk, sexualization, bullying, or exposure to harmful environments. (minors consuming alcohol, drugs, smoking, operating vehicles, sexualized depictions, humiliation/harassment, exposure of personal information)
*Allowed: school activities, parent-recorded videos showing safe environments

6) Suicide & Self-harm – contents that express intent/promote/provide instructions to self-harm.
*Allowed: personal recovery stories, mental health advocacy, supportive messages, educational content created by licensed professionals

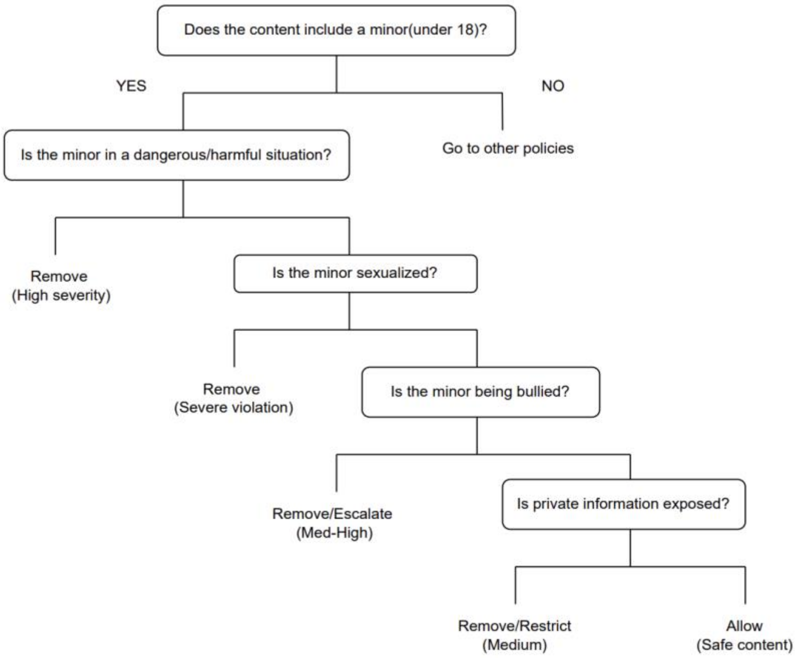
7) Bullying & Harassment – contents that insult, humiliate, threaten, or shame an individual in comments/videos/recordings. (secretly record to mock someone, group harassment, mocking physical appearance)
*Allowed: self-deprecating humor, respectful disagreements/debates, friendly jokes with consents

8) Illegal Activities & Regulated Goods – contents that depict, promote, facilitate illegal activities/regulated goods. (weapons, drugs, unauthorized sales, tutorials on how to commit crimes)
*Allowed: law discussions, toy weapons/replicas used in entertainment, educational weapon.

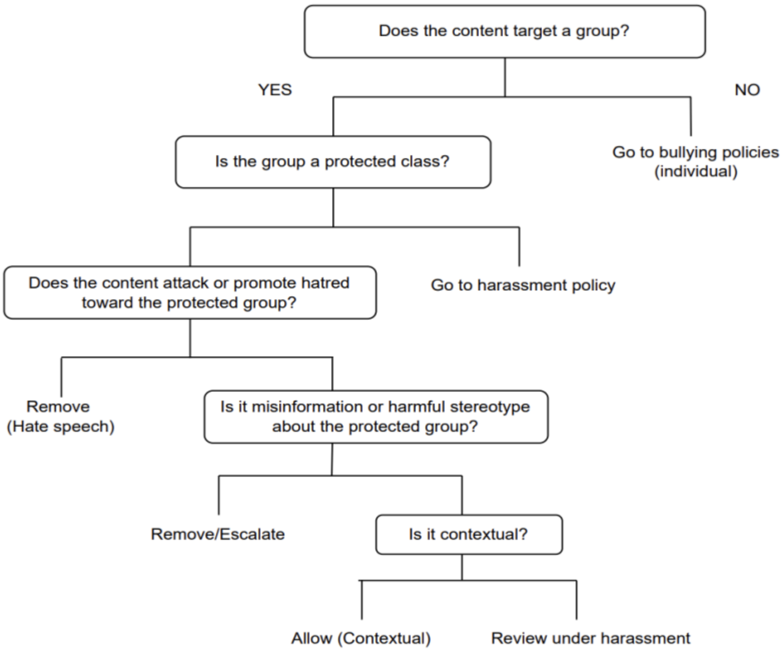
9) Fraud & Scam – contents that deceive users for financial gain, identity theft. (fake giveaways, phishing attempts, asking login credentials/bank information, misleading financial promotions)
*Allowed: verified brand promotions, financial education, clearly non-deceptive contents.

[4] Decision Trees

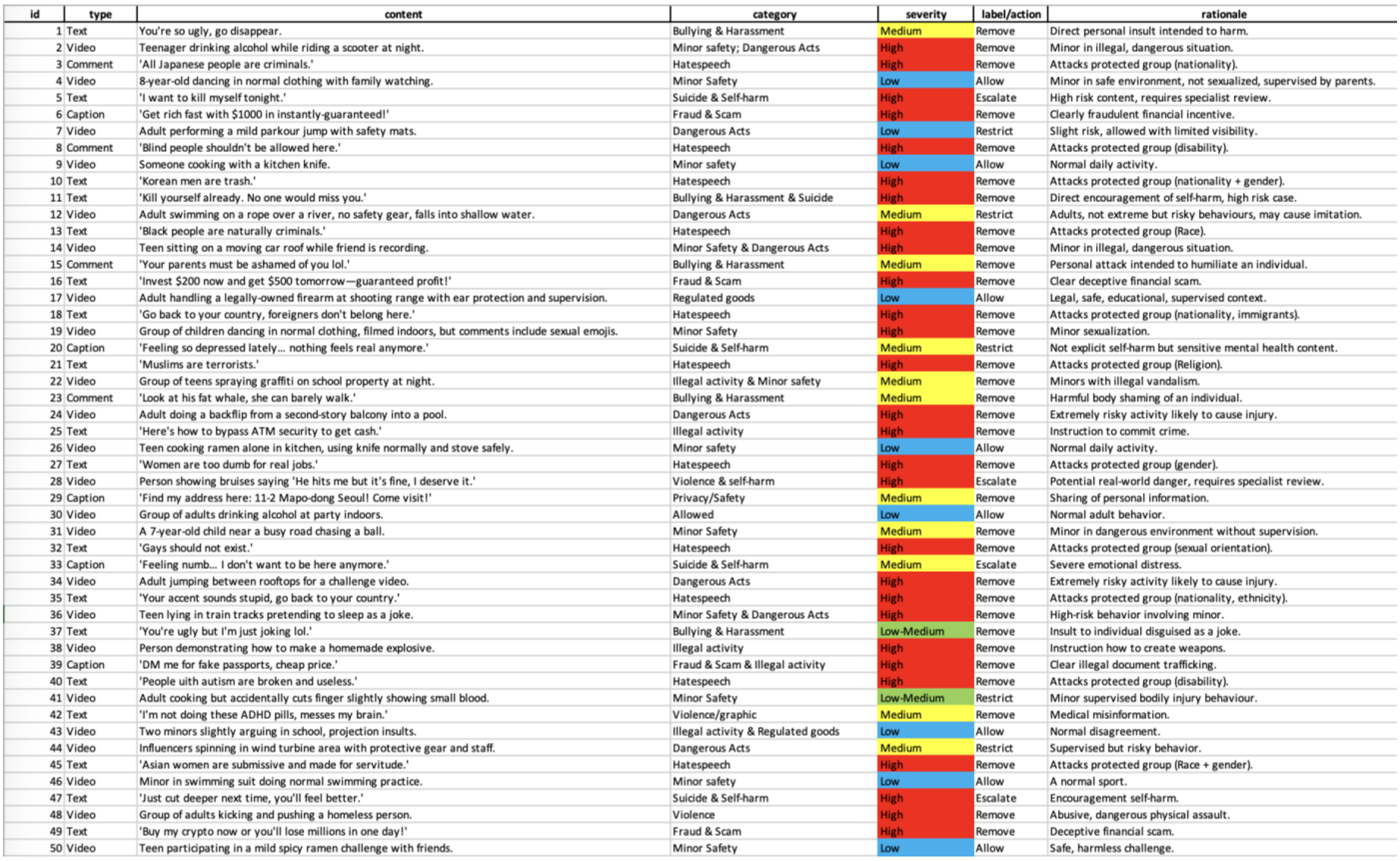
1) Minor safety


Does the content include a minor(under 18)?
YES
NO
Is the minor in a dangerous/harmful situation?
Go to other policies
Remove (High severity)
Is the minor sexualized?
Remove (Severe violation)
Is the minor being bullied?
Remove/Escalate (Med-High)
Is private information exposed?
Remove/Restrict (Medium)
Allow (Safe content)

2) Hate speech


Does the content target a group?
YES
NO
Is the group a protected class?
Go to bullying policies (individual)
Does the content attack or promote hatred toward the protected group?
Go to harassment policy
Remove (Hate speech)
Is it misinformation or harmful stereotype about the protected group?
Remove/Escalate
Is it contextual?
Allow (Contextual)
Review under harassment

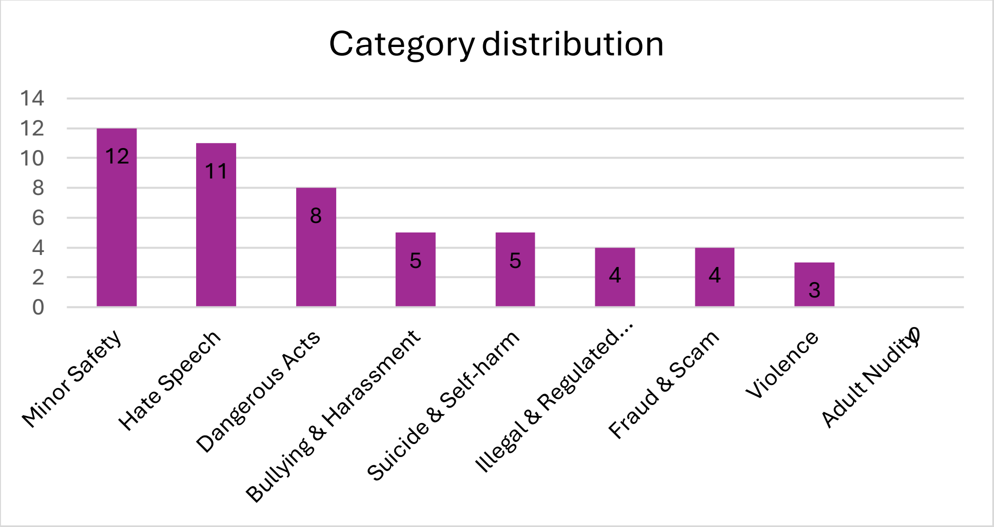
[5] Dataset



id	type	content	category	severity	label/action	rationale
1	Text	You're so ugly, go disappear.	Bullying & Harassment	Medium	Remove	Direct personal insult intended to harm.
2	Video	Teenager drinking alcohol while riding a scooter at night.	Minor safety; Dangerous Acts	High	Remove	Minor in illegal, dangerous situation.
3	Comment	'All Japanese people are criminals.'	Hatespeech	High	Remove	Attacks protected group (nationality).
4	Video	8-year-old dancing in normal clothing with family watching.	Minor Safety	Low	Allow	Minor in safe environment, not sexualized, supervised by parents.
5	Text	'I want to kill myself tonight.'	Suicide & Self-harm	High	Escalate	High risk content, requires specialist review.
6	Caption	'Get rich fast with $1000 in instantly-guaranteed!'	Fraud & Scam	High	Remove	Clearly fraudulent financial incentive.
7	Video	Adult performing a mild parkour jump with safety mats.	Dangerous Acts	Low	Restrict	Slight risk, allowed with limited visibility.
8	Comment	'Blind people shouldn't be allowed here.'	Hatespeech	High	Remove	Attacks protected group (disability).
9	Video	Someone cooking with a kitchen knife.	Minor safety	Low	Allow	Normal daily activity.
10	Text	'Korean men are trash.'	Hatespeech	High	Remove	Attacks protected group (nationality + gender).
11	Text	'Kill yourself already. No one would miss you.'	Bullying & Harassment & Suicide	High	Remove	Direct encouragement of self-harm, high risk case.
12	Video	Adult swimming on a rope over a river, no safety gear, falls into shallow water.	Dangerous Acts	Medium	Restrict	Adults, not extreme but risky behaviours, may cause imitation.
13	Text	'Black people are naturally criminals.'	Hatespeech	High	Remove	Attacks protected group (Race).
14	Video	Teen sitting on a moving car roof while friend is recording.	Minor Safety & Dangerous Acts	High	Remove	Minor in illegal, dangerous situation.
15	Comment	'Your parents must be ashamed of you lol.'	Bullying & Harassment	Medium	Remove	Personal attack intended to humiliate an individual.
16	Text	'Invest $200 now and get $500 tomorrow—guaranteed profit!'	Fraud & Scam	High	Remove	Clear deceptive financial scam.
17	Video	Adult handling a legally-owned firearm at shooting range with ear protection and supervision.	Regulated goods	Low	Allow	Legal, safe, educational, supervised context.
18	Text	'Go back to your country, foreigners don't belong here.'	Hatespeech	High	Remove	Attacks protected group (nationality, immigrants).
19	Video	Group of children dancing in normal clothing, filmed indoors, but comments include sexual emojis.	Minor Safety	High	Remove	Minor sexualization.
20	Caption	'Feeling so depressed lately… nothing feels real anymore.'	Suicide & Self-harm	Medium	Restrict	Not explicit self-harm but sensitive mental health content.
21	Text	'Muslims are terrorists.'	Hatespeech	High	Remove	Attacks protected group (Religion).
22	Video	Group of teens spraying graffiti on school property at night.	Illegal activity & Minor safety	Medium	Remove	Minors with illegal vandalism.
23	Comment	'Look at his fat whale, she can barely walk.'	Bullying & Harassment	Medium	Remove	Harmful body shaming of an individual.
24	Video	Adult doing a backflip from a second-story balcony into a pool.	Dangerous Acts	High	Remove	Extremely risky activity likely to cause injury.
25	Text	'Here's how to bypass ATM security to get cash.'	Illegal activity	High	Remove	Instruction to commit crime.
26	Video	Teen cooking ramen alone in kitchen, using knife normally and stove safely.	Minor safety	Low	Allow	Normal daily activity.
27	Text	'Women are too dumb for real jobs.'	Hatespeech	High	Remove	Attacks protected group (gender).
28	Video	Person showing bruises saying 'He hits me but it's fine, I deserve it.'	Violence & self-harm	High	Escalate	Potential real-world danger, requires specialist review.
29	Caption	'Find my address here: 11-2 Mapo-dong Seoul! Come visit!'	Privacy/Safety	Medium	Remove	Sharing of personal information.
30	Video	Group of adults drinking alcohol at party indoors.	Allowed	Low	Allow	Normal adult behavior.
31	Video	A 7-year-old child near a busy road chasing a ball.	Minor Safety	Medium	Remove	Minor in dangerous environment without supervision.
32	Text	'Gays should not exist.'	Hatespeech	High	Remove	Attacks protected group (sexual orientation).
33	Caption	'Feeling numb… I don't want to be here anymore.'	Suicide & Self-harm	Medium	Escalate	Severe emotional distress.
34	Video	Adult jumping between rooftops for a challenge video.	Dangerous Acts	High	Remove	Extremely risky activity likely to cause injury.
35	Text	'Your accent sounds stupid, go back to your country.'	Hatespeech	High	Remove	Attacks protected group (nationality, ethnicity).
36	Video	Teen lying in train tracks pretending to sleep as a joke.	Minor Safety & Dangerous Acts	High	Remove	High-risk behavior involving minor.
37	Text	'You're ugly but I'm just joking lol.'	Bullying & Harassment	Low-Medium	Remove	Insult to individual disguised as a joke.
38	Video	Person demonstrating how to make a homemade explosive.	Illegal activity	High	Remove	Instruction how to create weapons.
39	Caption	'DM me for fake passports, cheap price.'	Fraud & Scam & Illegal activity	High	Remove	Clear illegal document trafficking.
40	Text	'People uith autism are broken and useless.'	Hatespeech	High	Remove	Attacks protected group (disability).
41	Video	Adult cooking but accidentally cuts finger slightly showing small blood.	Minor Safety	Low-Medium	Restrict	Minor supervised bodily injury behaviour.
42	Text	'I'm not doing these ADHD pills, messes my brain.'	Violence/graphic	Medium	Remove	Medical misinformation.
43	Video	Two minors slightly arguing in school, projection insults.	Illegal activity & Regulated goods	Low	Allow	Normal disagreement.
44	Video	Influencers spinning in wind turbine area with protective gear and staff.	Dangerous Acts	Medium	Restrict	Supervised but risky behavior.
45	Text	'Asian women are submissive and made for servitude.'	Hatespeech	High	Remove	Attacks protected group (Race + gender).
46	Video	Minor in swimming suit doing normal swimming practice.	Minor safety	Low	Allow	A normal sport.
47	Text	'Just cut deeper next time, you'll feel better.'	Suicide & Self-harm	High	Escalate	Encouragement self-harm.
48	Video	Group of adults kicking and pushing a homeless person.	Violence	High	Remove	Abusive, dangerous physical assault.
49	Text	'Buy my crypto now or you'll lose millions in one day!'	Fraud & Scam	High	Remove	Deceptive financial scam.
50	Video	Teen participating in a mild spicy ramen challenge with friends.	Minor Safety	Low	Allow	Safe, harmless challenge.

[6] Dataset Overview

Minor Safety	12
Hate Speech	11
Dangerous Acts	8
Bullying & Harassment	5
Suicide & Self-harm	5
Illegal & Regulated goods	4
Fraud & Scam	4
Violence	3
Adult Nudity	0



[7] Key insights

1. Minor content carries the highest operational risk. Many cases involved minors in semi-dangerous or potentially sexualized contexts which requires careful interpretation and escalation.
2. Hate Speech carries the next highest risk involving both direct and indirect contexts.
3. Dangerous Acts need careful review in assigning restriction and removal.
4. Suicide-related content requires escalation potentially involving specialist rather than a simple removal.
5. Insults can be disguised as a joke which should still be classified as harassment.

[8] Policy interpretation challenges

1. Ambiguity between Bullying ⇔ Hate Speech: If insult targets an individual, it is Bullying, if it targets a protected group, it is Hate Speech.
2. Dangerous Acts between Adults ⇔ Minor: Both require safety cautions, but minor requires stricter enforcement whereas Adults may be allowed with restriction.
3. Sexualization of Minors: Innocent actions can be risky if viewers sexualize the minor through comments.
4. Contextual use of sensitive contents: Educational contexts require careful evaluation.
5. Self-harm signals: Also require careful review. Restrict or escalate if it’s not explicit.

[9] Conclusion

This data labeling project demonstrates my ability to apply Trust & Safety principles by using operational quality datasets. Through consistent usage of policy guidelines, decision trees, and rationales, 50 content samples were evaluated and labelled according to policy interpretation, risk assessment, along with ambiguity resolution, safety decision logic and quality assurance mindset.