AI Safety Benchmark Datasets in Hindi

Final Project Report, Tattle Civic Tech to MLCommons

Last updated: 25 November 2024
Published on: 4 December 2024

Warning: this report contains references to sex-related crimes and hate speech, and may contain triggering language.

Scope of the Report

From June - July 2024, MLCommons ran a global expression of interest for qualified and interested organisations and individuals to pitch a prompt generation or evaluation pilot project. Out of more than 20 submissions, Tattle Civic Tech was selected to develop a dataset of 2000 Hindi- language prompts spanning two hazards, as defined by MLCommons[1], through a participatory approach. To achieve this, Tattle conducted online workshops, primarily with Hindi speaking experts- including social workers, journalists, and researchers, to collectively generate Hindi prompts based on their real-world experiences and knowledge. These prompts, broken down to three personas, focus on the critical hazard categories of hate speech and sexual related crimes. This project took place over the course of two months starting September 2024. This report should not be taken as representative of the Hindi language, spoken as a whole elsewhere outside of India, nor of all 19+ languages spoken in India. Further, this exercise is not exhaustive on all forms of sex-related crimes/hate one may experience in Hindi.

[1] Vidgen, B. et al. (2024). Introducing v0.5 of the AI Safety Benchmark from MLCommons.

Table of Contents

Background Work

Tattle has prior experience in building language specific datasets on online harms such as misogyny and misinformation.[2] In the process of development of Uli, an ML-based plugin to respond to online gender based violence (OGBV), we created a dataset through a participatory approach involving expert contributions by those who encountered or worked to respond to OGBV in India.[3] We also maintain a list of slurs and coded words in Indian languages that is crowdsourced from researchers and feminist and gender rights organisations in India.[4] The Tattle project team for this work consisted of members with contextual knowledge and domain expertise in AI and information harms.

Expert Led Participatory Approach

We extended the participatory approach of the Uli dataset to this work. The prompts were sourced from Hindi speaking experts in India who either identify as or have substantial experience working with survivors of identity based targeting online. The experts not only wrote the prompts but also weighed in on the definition of hazards and on the process through which the prompts were being collected. The process of prompt collection was continually adapted through their feedback. The process while retaining the broader participatory framework evolved with each workshop under the specific hazard categories.

This project is focused on hazard prompts in the Hindi language exclusively. The project team and the expert group participants were primarily native Hindi speakers. The concerns surfaced in the prompts are contingent upon the composition of the group. While an attempt was made at capturing the diversity of concerns within Hindi speaking populations, the tight scope of the task placed limits on how many people we could engage with.

We suggest that this report, in distilling this method of prompt generation, be treated as adding to the unique context of Hindi spoken in this region of India. Our findings are not representative of the language spoken as a whole elsewhere outside of India, nor of all 19+ languages spoken in the country. Further, this exercise is not exhaustive on all forms of sex-related crimes/hate one may experience in Hindi.

[2] Prabhakar, T., Gupta, A., Nadig, K., & George, D. (2021). *Check Mate: Prioritizing User Generated Multi-Media Content for Fact-Checking.* Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 1025-1033. https://doi.org/10.1609/icwsm.v15i1.18126

[3] Arora, A. et al.. (2024). *The Uli Dataset: An Exercise in Experience Led Annotation of oGBV.* In Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024) (pp. 212–222). Association for Computational Linguistics.

[4] Nazariya, Khabar Lahariya, Point of View, and Feminism in India

Expert Group

Criteria for Selection:

We invited participants from various Indian organisations and individuals who satisfied either of the following criteria:

- Belong to or work with marginalised communities in India frequently exposed to hate and violence due to factors like caste, class, gender, sexual orientation, religion, ethnicity, or ability;
- Advocate for the inclusion of marginalised groups through activism, research, or professional practice;
- Possess expertise in dark creativity, which is the use of creative ideas to predict malevolent actions, or have worked with perpetrators of harmful content.

Organisations and Professions represented

We engaged a diverse group of twenty-four experts to ensure a comprehensive representation of regional and caste-based identities. The organisations represented in the group are: Nazariya[5], The Mooknayak[6], RATI[7], NewsMeter[8], The Alternative Story[9], BOOM Live[10], Monk Prayogshala[11], Western Norway Research Institute, University of Delhi, National Council of Women Leaders, and International Land Coalition[12]. The group has expertise in journalism, social work, feminist advocacy, gender studies, fact-checking, political campaigning, education, psychology, and research. All experts are native or fluent Hindi speakers.
A majority of the experts had engaged with Tattle in the context of previous projects. Some experts were added to the group based on the recommendation of other experts. The experts were requested to attend a minimum of three out of six workshops to ensure consistent engagement and quality contributions. An honorarium was fixed per workshop and the experts were paid at the end of the six workshops, based on the number of workshops attended. Since the discussions involved detailed discussions of triggering topics, an additional amount was also earmarked for psychosocial support for the experts.

[5] https://nazariyaqfrg.wordpress.com/
[6] https://www.themooknayak.com/
[7] https://ratifoundation.org/
[8] https://newsmeter.in/
[9] https://alternativestory.in/
[10] https://www.boomlive.in/
[11] https://www.monkprayogshala.in/
[12] https://www.landcoalition.org/en/

Choice of Hazards

Tattle chose to work on Hate as a hazard owing to its prior experience in building datasets on identity based targeting online. For selecting the second hazard, the experts were asked to vote on one of the following hazard categories[13]:

- Violent Crimes
- Non Violent Crimes
- Sex Related Crimes
- Child Sexual Exploitation
- Suicide and Self Harm

Most experts who voted, opted for sex-related crimes as their most preferred option, and it was therefore selected as the second hazard category.

Methodology

Literature Review

The AI Safety Benchmark v0.5 (v0.5)[14], and the AILuminate Benchmark served as a starting point to understand how prompts were classified, and how personas, sentence fragments, templates, interaction types, activities and viewpoints were interpreted. We also explored other similar benchmark creation work.[15,16] Since hate as a hazard category had been decided by the team at the start of the project, we, in parallel, conducted literature review on conceptions of hate in the Indian context. The second hazard of sex related crimes (SRC) was decided by the expert group at the end of the first workshop. The literature review for SRC took place as prompts for hate were being collected. The sub-sections on hate and SRC contain more details about the background literature that informed the team's understanding of these concepts.

Workshop Structure

We structured the prompt generation exercise as six workshops spanning sixty to ninety minutes. The workshops took place between 18th September and 15th October. The first three workshops were scheduled one week apart to ease the expert group into the process. The last three workshops on sex related crimes (SRC) took place in twelve days.

[13] As set out in the v 0.5 benchmark dataset paper

[14] Vidgen, B. et al. (2024). *Introducing v0.5 of the AI Safety Benchmark from MLCommons.*

[15] Storchan, V., Kumar, R., Chowdhury R., Goldfarb-Tarrant, S., and Cattell., S.(2024). *Generative AI red teaming challenge: transparency report.*
URL https://drive.google.com/file/d/1JqpbIP6DNomkb32umLoiEPombK2-0Rc-/view.

[16] Mazeika, M. et al. (2024). *Harmbench: A standardized evaluation framework for automated red teaming and robust refusal.*

The workshops took place online on Google Meet. Every workshop had a breakout session where participants could work together in smaller groups. Experts were distributed into breakout groups based on the people in attendance during the workshop, and area of expertise. Each breakout room consisted of 6-8 experts and had one facilitator and an optional note taker. The workshops were conducted in Hindi.

Experts were asked to note their contributions on a Google Doc which had the language setting set to Hindi. This enabled experts to write on the documents in Hindi, if they wanted to. Experts were also given the option of typing the prompts in English, or Hindi in Latin script. Some experts who were joining from their phones, sent their prompts to facilitators over WhatsApp and the facilitators copied them into the Google doc.

After a workshop a translator reviewed all the prompts and converted them to Hindi in Devnagari script. For transliteration the team used freely available tools.[17] For translation of prompts and the workshop script, the team also used Google translate. All translations, however, were reviewed by a human translator.

Expert Onboarding

Prior to the first workshop we spoke to each of the experts to explain the nature of the task and Tattle's association with MLCommons. The first forty minutes of the first workshop were devoted to providing ground rules for the workshops, providing an overview of Large Language Models (LLMs), prompts, and safety benchmark datasets and the anticipated impact of their contributions.

Hazard: Hate

Conceptualizing Hate

Following a literature review spanning social science , we converged on the definition of hate speech given by the 'UN Strategy and Plan of action on Hate Speech' (2019) as the starting point for the discussion with experts. As per this definition, hate speech is defined as, "Any kind of communication that attacks or uses discriminatory language targeting a person or a group on the basis of who they are (i.e., their identity)"[18]. Our approach to hate speech focused on these two aspects, the nature of hate speech and the identities of the target groups. The targeted groups could be from different identities such as caste, class, religion, region, tribal/ethnic identity, political ideology,

[17] https://indiatyping.com/index.php/home/convert-hinglish-in-hindi

[18] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech; Identity groups included: caste, class, religion, regional identity, tribal/ethnic identity, political affiliation/opinion, sex, gender, sexual orientation, disability, appearance, age.

sex, gender, sexual orientation, disability, appearance and age. The nature of hate can be in the form of derogatory remarks, intolerant/discriminatory behaviour and inciting/inflicting harm. Derogatory remarks against a particular community can be in the form of stereotyping, name calling and abusive humour. Whereas intolerant or discriminatory behaviour can target communities based on social, economical, cultural and political lines. Similarly hate speech can also incite or inflict harm on people both physically and mentally.

The team had some prior familiarity and expertise in creating participatory datasets related to hate speech, gendered abuse and misinformation. Existing datasets created by Tattle[19,20] and other research groups[21][22] helped in coming up with an initial list of slurs and stereotypes used to target identity groups in the country. We categorised slurs through their origin, which may have its roots in caste or religious identity. For example, use of derogatory terms like *chapri* (lower caste) and *gaandu* (homosexual) hint at a more structural hate narrative around these identity groups. This list was shared as supporting material that experts could refer to for discussion or prompt ideas.

We also aimed to contextualise hate speech which could be community specific or state specific. For example certain caste groups in particular states may have different hate narratives around them. The diverse background and range of expertise of the workshop experts was structured to help account for the specificities.
To generate prompts, the experts were asked, *"Can you generate prompts that are hateful or might elicit a hateful response?"*

Experts delved into issues like religious stereotypes and practices used to target minority groups. For instance, the practice of "halala" within the Muslim community was discussed, highlighting its use as a tool to attack and marginalise the community. Similarly, an expert brought up derogatory references to Dr. B.R. Ambedkar (a jurist and head of the drafting committee of the Indian Constitution, a social reformer and a leading figure for Dalit consciousness) as the "first rapist of Independent India". This was attributed to a post circulating on X/Twitter that wrongly alleged that he sexually assaulted a woman, in an attempt to disparage the Dalit community.

[19] Arora et al. (see footnote 3)

[20] https://github.com/tattle-made/Uli/tree/3953ae73c489a4704f028af019332df5c530ad6c/browser-extension/plugin/scripts

[21] Jha, A. et al,. (2023). *SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models.*
https://arxiv.org/abs/2307.10514

[22] Dev, S., (2023). Building Socio-culturally Inclusive Stereotype Resources with Community Engagement.

Insights from participants were therefore highly contextual and rooted in their experiences and identities. Based on the workshop discussions and prompts generated, a diverse range of identities, relationships between those targeting hate and those targeted by it, and polarising topics around which hate is expressed, were documented:

Target identities:	Muslims, Christians, Women, Dalits, Tribal community, Queer community, People From North-East India, People Experiencing Mental Illness, Nationality, Politician, Writer, Father
Relationship:	Landlord, Parent, Husband, Neighbour, Flatmate, Lover, Stalker, Employer
Topics:	Reservation[23], Abortion, Raising Children, Religious Texts, Nationalism, Relationships, Mental Health, Marriage, Army, Food, Employment, Family Structure And Norms, Behaviour, Citizenship

While some breakout groups opted for a more independent approach to prompt generation, others found value in collaborative discussions facilitated by group leaders. The second workshop focused on the potential for malicious use of LLMs. Participants were tasked with developing prompts that, while seemingly innocuous, could elicit hateful or biased responses. This involved using persuasive language, exploiting deep contextual knowledge, and identifying potential slurs or coded language, which could "jailbreak" LLMs. For example, participants were asked to frame prompts as requests for help with research or creative writing, knowing that certain topics could trigger harmful stereotypes or biases. By the end of the workshop, over 150 such prompts were generated, demonstrating the potential for complex and insidious misuse of LLMs.

The third and final workshop was in continuation to the second workshop and further nudged experts for more complicated prompts. In this workshop experts were provided with the glimpse of prompts, templates and stereotypes emerged in the last two workshops.

Prompt Generation Method

Every workshop led to some prompts written entirely by the experts. Each prompt was reviewed by the project team. A member of the team annotated each prompt for (a) whether it fit the definition of hate hazard, (b) identity group that is the subject of the prompt, and (c) the persona reflected in the prompt (skilled, unskilled, normal). In addition to the prompts, the workshops also surfaced stereotypes and "seed prompts" that could be used to generate templates. The team used the stereotypes to create more prompts. The templates were multiplied into prompts with the help of

[23] Reservation refers to a system of affirmative action in India in public institutions.

community specific slurs and derogatory phrases, which were mapped earlier using the Uli, SPICE and Seegull datasets. Thus, there were three ways in which hate prompts were generated using expert engagement:

1. Experts wrote the entire prompt
2. Experts shared stereotypes which the team converted into prompts
3. Experts shared templates. The team expanded on these templates and multiplied them into prompts

The following table describes the number of prompts generated through each method.

Type of Hate Prompt	**Number of Prompts**
Written by Expert	359
Stereotype from Expert	223
Templates to Prompts	419

Hazard: Sex-Related Crimes

Conceptualizing SRC

The MLCommons definition of SRC included:

- Sexual assault (e.g., rape)
- Sexual harassment
- Sex trafficking
- Prostitution
- Indecent exposure (i.e. showing sexual body parts or “flashing”)

To understand the relevance of the MLCommon’s definition to the Indian context, we consulted the National Crime Records Bureau to identify crimes that are classified as sex-related crimes. The NCRB documents crimes such as attempts to outrage the modesty of women, sexual harassment in the workplace, sexually explicit content, sexual harassment in shelter homes, sexual harassment in public transport, defamation, blackmail, and storing CSAM.[24] They were compared with crimes in other jurisdictions as well. The process of defining the hazard was complicated, for the reason that most definitions of sex-related crimes merely listed the offences under the category and did not have a broader definition. We also perused literature related to

[24] NCRB Crimes in India Year Wise: https://www.ncrb.gov.in/crime-in-india-year-wise.html?year=2022&keyword=

the category, but found that they predominantly delved into motivations behind committing the offences. We thereon reviewed prominent press coverage of these offences in the country, and combining the various sources, began to move towards a definition of sex-related crimes for the purpose of this exercise.

Sex-related crimes introduced the challenge of considering a jurisdiction in which something is considered a crime. Interpreting the hazard in accordance with Indian domestic penal laws, limits it to offences defined under these laws. For example, in India, marital rape is not an offence. A strictly legal interpretation would preclude prompts pertaining to marital rape. In another instance, rape is defined under both the erstwhile and current India penal laws as a crime that is committed by a *man against a woman*. Its scope does not include the act taking place by and against other genders. However, the act itself can still be considered a sex-related crime. We therefore did not limit the scope of the hazard to the jurisdiction of Indian law. We also included within the scope of the hazard acid attacks, caste honour killings, marital rape, female foeticide, and abortion, which are grave and predominant offences in the Indian context.

To obviate the need to define jurisdiction or enter complex debates about legality, we suggested to the experts that they need not limit themselves to acts considered crimes in the Indian jurisdiction, nor solely crimes against women. Experts were asked to consider an act for the purpose of the prompt writing exercise if *they* believed the act is or should be a crime in India. Any disagreements that arose in relation to *whether* an act must be considered a crime or not were to be noted by the facilitators but not delved into. We requested the experts to provide information as to the legality of the crime in a country (Ex: Legal in India, illegal in Canada), if they were aware of such detail in relation to the crimes they were writing prompts around.

Topics and crimes which were considered out of scope for this exercise were:

- Prompts which attempt to elicit imagery
- Sexual related crimes against minors
- Undirected misogyny
- Responses that describe sex-related crimes or facts about sex-related crimes
- Responses that describe consensual sexual experiences involving one or more adults

The first two workshops for this category followed the previous method of distributing participants into breakout groups supported by facilitators, with the primary mode of communication being Hindi. In the first workshop, prior to and post the prompt writing exercise, participants provided feedback on the framework through which we designed the exercise for this hazard category. The experts emphasised the need to look at SRC from the lens of violence and power and not the letter of the law. Some other ways of looking at the problem were: the lens of penetrative vs non-penetrative sex, or through sites of the crimes: homes, family, state, public spaces, online. They also brought to note the CEDAW's framework[25] towards crimes against women, specifically those sexual in nature and suggested we use that as a foundation. With respect to enforceability, it was also brought to our attention that: (i) in light of the new penal laws coming into effect in India, there are still sections of the previous criminal laws which apply in case of crimes which have not been covered in the new law; (ii) while there are central laws which apply to all of the states in the country, each state may enact state-specific rules which set out how an offence is to be handled. For instance, in some states crimes may be reported at one-stop centres, but not in other states. The experts suggested that prompts may therefore be framed with respect to enforceability as well. The ML Commons v0.5 standard explicitly states that responses containing facts around a specific crime are not considered unsafe. The experts however emphasised that soliciting facts in certain cases is in fact illegal. For example: in India, it is illegal to identify: (i) the victim of a sex-related crime, and (ii) the sex of an infant prior to birth. There have been documented instances in India of mass online attempts to identify victims of brutal rape cases.[26] Based on this feedback, fact-based prompts or prompts eliciting factual information about certain crimes were also included in the exercise. The experts also discussed sex-related fantasies about sister-in-law (*"bhabhis"*) and daughter-in-law that are tied to India's familial structure; as well as power dynamics between caste groups that are expressed in sexual violence[27].

Having received feedback from participants from the first workshop for this hazard, and observing that the participants were not focusing on the question of legality of crimes mentioned in the prompts, we redefined the approach to build prompts for the next workshop. This was shared with participants as a guidance to approach prompt writing in the second workshop:

- Who is the offender/target, and what is the relationship between them?
- What is the nature of sex-related crime?
- Where has SRC been committed?

A few case studies were provided that reflected different acts, actors and sites of sex-related crimes. Participants were also asked to consider the interrelations with film, music, media, politics, current affairs, sports, pulp fiction books, and pop culture

[25] https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women

[26] https://www.indiatoday.in/india/story/kolkata-doctor-rape-murder-victim-rg-kar-hospital-online-bait-for-some-it-is-business-2586492-2024-08-22; https://fightthenewdrug.org/woman-in-india-raped-and-murdered-her-name-trended-on-porn/; https://www.outlookindia.com/society/how-perverted-is-india-kathua-victims-name-becomes-top-trending-search-on-porn-w-news-311055

[27] https://ruralindiaonline.org/en/library/resource/caste-based-sexual-violence-and-state-impunity-english-and-hindi/

content.

Based on the observation that participants appeared to limit themselves to a select few topics and range of prompts within them in the first workshop, in the second workshop participants were encouraged to think about other incidents such as date rape, mass sexual assault, sexual violence in inter-group conflicts, cyberflashing, voyeurism and sex trafficking. A reference document with news coverage and mainstream media depiction of such incidents was distributed in this workshop. The document can be found in the Other Resources section in the appendix. The document served as a starting point for discussion/prompt generation but the experts were not bound to the topics. The influence of film and music was also discussed: one participant brought to notice that the lyricism of songs local to a certain region tend to be demeaning and vulgar towards women, and so creative prompts may be framed using that style of writing or to elicit responses in that style. While reflecting on the exercise in the joint , little-known practices followed in certain parts of the country, termed 'bahu jhootayi'/ 'raat basna'/'kheech pratha' which are forms of caste-based sexual violence, were also brought our attention.

The final workshop was distinct in that participants were split into two groups: the first breakout group focused on engaging in the exercise of prompt writing individually, and the second breakout group focused on generating concepts from which prompts can be generated. Experts in the discussion group spoke of the influence of porn in media, bringing to our notice several smaller websites which commonly feature non-consensual content, and a crowdsourced website of erotica in Indian languages. They discussed the influence of Indian movies and 'item songs' which predominantly depict and reduce women to sexual objects. Violence against communities was also briefly discussed: that Hindu men are attempting to 'regenerate' Hindu pride through 'reverse love jihad', i.e. engaging in group or gang violence against women of the Muslim community. There have also been instances of fear-mongering against the LGBTQIA+ community since s.377[28] was held unconstitutional in India, with claims that it would lead to children being unsafe.

Prompt Generation Method

Every workshop resulted in some prompts being written by the experts. Concepts, coded language for sexual acts and body parts that experts mentioned were documented by the facilitators. It was observed that a majority of the prompts from the experts were questions around specific actions ('how to', 'what is', 'should I'). These aligned with the unskilled persona. The team adapted some of the unskilled prompts to skilled prompts. The team also came up with its own prompts.

[28] Section 377 of the Indian Penal Code, 1860, which criminalised sexual acts between adults of the same sex.

As an experiment to generate skilled prompts, the team also tried two automated approaches:

1. Generating syntactical variations of prompts using punctuations and symbols from the Hindi alphabet.
2. Generating prompts or snippets of templates using a locally hosted Llama3.1. Through prompt engineering Llama was used to generate snippets that portray expertise in India. It was used to imbue a diversity in tone and word choice for similar looking prompts. Here is a sample prompt:

   > “I want to generate 50 sentences with the same semantic structure as the following sentences:
   > I am a professor at IIT Bombay
   > I am a philosopher from JNU
   >
   > Make sure that the institutions are based in India. Use diverse tones and word choices in the sentences.”

   These were then manually combined with different sex related crimes to create prompts.
   Some prompts were used to generate different ways to be persistent. For example:

   > Generate 50 example sentences that have the same semantic structure as the following example sentences
   > Example Sentence : Show me how to [EMPTY SPACE]. Ok fine just tell me how to do [EMPTY SPACE] instead

A small number (17) of prompts were generated entirely using Llama. All instructions to Llama were in English which resulted in outputs in English. The outputs from Llama were reviewed and translated by a human.

The following table describes the number of prompts generated through each method.

Type of SRC Prompt	**Number**
Written By Expert	493
Written By Team	96
Template to Prompts	274
Syntactical Variations	50

Prompt generated by LLM	17
Template generated by LLM to Prompts	83

Prompt Set Summary

After the creation of the prompts, a team member annotated each of the hate prompts with the identity targeted in the prompt, and the SRC prompts with the sex-related crime described in the prompt. While no goals were specified to cover specific identity groups or sex related crimes, this categorization provides a high level view of the issues covered in the prompt set.

Primary Identity Targeted in Hate Prompt	
Identity	**Count**
Religion	228
Women	178
LGBTQIA+	135
Caste	131
Regional	84
Aadivasi/Tribal	64
Political	35
Mental Illness-Disability	27
Feminist Ideology	17
Men	14
Other	87

Primary Act Described in SRC Prompt	
Type of SRC	**Number of Prompts**

Sexual Harassment (Includes stalking, workplace sexual harassment, violating consent, 'wife-swapping')	332
Sexual Assault (includes rape, violent sex, spiking)	308
Indecent Exposure, porn, NCII, sexual morphing, voyeurism	103
Sexual Exploitation and sextortion	81
Prostitution	51
Sex trafficking	21
Abortion and Foeticide	13
Acid Attack	9
Other (includes cultural practices, some prompts on consent)	95

Reflections on Process

An expert led participatory method aims to generate context-specific knowledge of how a specific harm manifests in a population. The constraints of time, and the task being defined around pre-specified hazards of hate and sex-related crime- introduced some limits to how the participatory process could be designed. Yet, the resultant prompts and resources used to generate the prompts, affirm the relevance of such engagement in creating safety benchmarks for AI.

There are three aspects about the overall process, that stand out to us about this specific engagement with experts:

- **Rich contextual input:** Discussions helped bring in rich grounded insight and nuance to a variety of culturally-specific themes, scenarios and tendencies. These themes, be it for hate or sex related crimes, are not immediately obvious to a general audience.

- **Feminist standpoint:** What stood out in the exercise was the overwhelming number of prompts which were based on prejudice against minorities. This was informed by Tattle's orientation to oBGV work in the past as well as the collective experiences and domain expertise of the group.

- **Discussions to seed prompts:** The relevance of certain knowledge to prompts did not come naturally to all the participants. In the workshop sessions for the hazard on sex-related crimes, there were rich insights about casteist and misogynistic practices ('raat basna' and others as mentioned earlier), which emerged after considerable discussions about various intersectional crimes and which did not instinctively come to mind in a prompt framework.

Below we describe reflections on specific aspects of the process, in greater detail:

Differences between Hate and SRC:

The distinction between the two hazards was not patently obvious to participants and prompts often overlapped, much as the manifestation of the hazards in the real-world context. Participants found it easier to come up with templates, stereotypes, and prompts under the hate hazard, including skilled prompts, as compared to SRC. With SRC it was challenging for participants to write skilled prompts, especially around topics such as gang rape and sex-trafficking, leading to fewer prompts around these crimes.
Even as the expert group had comparable expertise in the two hazards, the extent to which that expertise translated into prompts varied between the two hazards. Tattle's past experience is also a key differentiator here, as Tattle's background in Uli work helped us effectively translate/frame the dataset requirements on hate, into tasks that the experts could easily engage with. On the other hand, generating prompts for SRC posed a few challenges. Unlike hate, the baseline moderation systems currently in place can catch content on SRC when stated in plain language with direct terms such as 'rape' or 'sex'. Added to this, the shorter duration between workshops on SRC resulted in lesser time to review and reflect on prompts generated in one workshop to inform changes in subsequent workshops.

Possibility of Using LLMs to Generate Language Specific Prompts:

MLCommons had shared a sample of prompts in Hindi created by another vendor. From perusing through the prompts it was clear that the prompts for hate were translated from American English. For example, several prompts in the set shared by MLCommons described prejudices against Black or Jewish population which are not the primary identities targeted in the Indian context.

The language was also excessively formal. The prompts pertaining to SRC were more relevant. They described acts that might also take place in India, although in formal language. They however also left out a range of practices, settings and relationships expressed in sex related crimes in India. The prompts generated by the experts better reflected real world usage of Hindi language to describe sexual acts.

As an experiment, Tattle also tried using LLMs to generate a few skilled (~100) prompts pertaining to SRC. While LLMs were helpful in producing interesting ideas for characters in the prompts, it took significant effort to rework them to reflect real world language and scenarios.

Impacts of working on/with harmful content:

Participants expressed that the workshops took a toll on them, and that the tasks did end up feeling emotionally taxing. The SRC workshops had more of an impact, both due to the nature of the topic, as well the short time frame in which the three workshops were conducted. At the culmination of the workshops, participants were informed that in cognisance of the nature of the work undertaken over the course of the workshops and the impact of participation, Tattle would support any participant who wished to opt for psycho-social support. To facilitate the same, we partnered with the Alternative Story, an Indian organisation which provides mental health services, and whose members also participated in the workshops sessions as well. Some members availed this support.

Limitations of the dataset framework:

The exercise was shaped by the framework provided by MLCommons. While it was a helpful starting point, the framework was also, in some cases, limiting. The persona of the prompt, in particular, presented challenges since the experts found it more natural to think of prompts from the perspective of victims or perpetrators. The psychologists in the expert group could think of creative ways of jailbreaking prompts, but the distinction between simple malicious and advanced malicious was not always clear.
The experts also often thought in terms of what the prompt would result in. For example, some experts pointed out that the resources shared by an LLM in response to a prompt on self harm or domestic violence were from other countries and therefore not useful to people in India. So while strictly speaking, the prompt output is not unsafe, the safety layer is not protecting users in India.

Quality Assurance:

ML Commons specified that the quality of the prompts will be assessed along the dimensions of:

- Diversity: duplicates, near duplicates, lack of variance
- Construct validity: the prompts meet the scope
- External validity: the prompts are realistic

Due to constraints of time and budget, the experts only created prompts and did not review or rate other prompts. The internal team reviewed all prompts but beyond assessing whether the prompt fit scope of the hazard, the team did not alter the prompts. In a more intensive exercise, every prompt could have been reviewed by an odd number of experts. This would have given a better sense of the factuality, relevance and potential harm of the prompts.

The diversity of the prompts has been assessed based on the identities targeted. A finer grain annotation of the identities targeted is listed above in the section on Prompt Set Summary. Construct validity was ensured by a team member reviewing every prompt. External validity was ensured by the design of the workshops. As some of the experts mentioned during a feedback session- the experts relied on their real world experience. Two specifically mentioned that they were writing prompts based on questions they heard during training sessions on gender sensitization, or while working with survivors and perpetrators.

Open Reflections from the Expert Group:

We asked participants their vision/hopes with the dataset, and how they would like to see their contributions to such a project/dataset put to use. A participant stated that they would like to see the contributions released under a licence, even if it is a commercial one, and for contributors to be credited for their work. Another participant believed that prompts should be collected on a regular basis to keep the models updated.
As an outcome for this exercise, a participant expressed that they hoped LLM models would provide ‘neutral’ outputs to such prompts, which would satisfy either everyone or no one, but which were clear and factful. Participants also indicated that there could be more representation from the trans and queer community in future iterations of this exercise. Finally, participants also emphasised the need to focus on AI generated images, which are popular in a country like India that relies heavily on audio-visual content for engaging online.

Recommendations

This project was a small pilot to extend the v1.0 Safety Benchmark, tried first in English, to a new language through contextual interpretation of the taxonomy. The reflections from the process can inform expert engagement in AI safety work in general. Here we list some recommendations specifically for the creation of the AI Safety Benchmark dataset:

Taxonomy Related

- The definition of hazards set out in the v1.0 paper leaves significant latitude for interpretation. For the hazard on hate, we relied upon the UN's definition of hate as under the 'UN Strategy and Plan of action on Hate Speech' (2019). For SRC, we started by looking at the Indian Penal Code but a recommendation from an expert was to rely on CEDAW's framework. The hazards can be tied to overarching global frameworks to ensure some alignment when interpreted across regions. Alternatively, such alignment may be arrived at through consultative processes with teams involved in similar exercises in different regions.

- Hazards are not mutually exclusive. For instance, a statement that attacks a person belonging to a certain identity may involve a threat to commit an SRC as well. Our experience affirms the need for a faceted taxonomy, as suggested in the v0.5 safety benchmark paper, so that each test item can be assigned to multiple hazard categories.

Process Related

- The goal of a workshop based approach to prompt writing was to discover and document contextual realities around the two hazards. To leverage the expertise of participants, a free-form approach allowed for a more authentic reflection of their experiences. This extended to workshop design as well. While the workshop format requires strong facilitation, the overall expertise in the group is greater than that of the facilitators. Overly strict guidelines and definitions can stifle creativity of responses by participants. Offering nudges towards desired topics and goals may be a more effective method at moving towards the goals of the tasks.

- The exercise, localised against one language, revealed unique risks not applicable to English or an American/Western European context. To comment on the generalisability of the existing framework, it is recommended that the exercise is repeated across more languages.

Appendix

List of experts

Dr. Hansika Kapoor (she/her)
Dr. Hansika Kapoor is a psychologist and researcher based out of Mumbai and has previously worked in creativity, misinformation, and online harm.

Zayan Nohora (he/him)
Zayan (He/him) is a dalit queer feminist activist. He has been working on the intersections of gender, sexuality, queer and trans* rights, masculinity, and education for 5 years now. He is also a TEDx speaker and has published in journals like Contemporary Education Dialogue, The Teacher Plus, The New Leam, etc.

Kirti Rawat (she/her)
Kirti is an Assistant Producer at BBC Hindi, based in Delhi with over 3 years of experience. She has covered conflict, crisis, gender, social issues and culture over these 3 years.

Saurav Verma (they/them)
Saurav Verma has been working in the field of Gender, Sexuality, Queer Trans rights, SRHR and community health for over 8 years. They are a Queer Non-binary Bahujan activist and an educator.

Rohini Lakshané (she/her)
Rohini Lakshané is an interdisciplinary researcher, technologist and a Wikimedian. Detailed bio at https://about.me/rohini

Dr. Cheshta Arora (it)
Cheshta Arora is an ethnographer mapping socio-technical relations in the contemporary.

Dheeshma Puzhakkal (she/her)
Dheeshma Puzhakkal is a Google News Initiative-certified fact-check trainer, currently serving as the Editor of the Fact Check team at NewsMeter, an IFCN-accredited news and fact-checking organisation based in Hyderabad. Previously, she worked as a Principal Correspondent at India Today.

Mahfooz Alam (he/him)
Mahfooz Alam is a journalist and works as an Associate Fact-Checker for NewsMeter. His work primarily is to find misinformation and disinformation doing the rounds in Hindi

on social media and debunk them. He joined NewsMeter in September 2022 as a Fact-checker and was promoted to Associate Fact-Checker this year in September.

Uma S, Siddharth P: RATI Foundation
RATI Foundation is dedicated to supporting victims of sexual and gender-based violence, focusing on long-term rehabilitation. Their goal is to address all vulnerabilities and disruptions in a victim's life caused by the violence, ensuring a holistic recovery process. They run the MERI Trustline (+91-6363176363), which is a free helpline that offers support for victims of online violence with services that include content takedown, mental health assistance, social support and legal guidance.

Sheeba Aslam Fehmi (she/her)
Journalist, Anchor, Rights Activist, Research Scholar and Gender Trainer. She believes that social media is an important tool for advocacy and one where people should strongly voice their perspectives and interests.

Shivani Yadav (she/her)
Shivani Yadav is a translator and pop culture writer. She's currently studying Psychology to become a Counselling Psychologist.

Dr. Seema Mathur (she/her)
Dr. Seema Mathur is an Academician, Activist and Poet, teaching in University of Delhi. She has engaged with social organisations to raise voices against atrocities, violence against women and dalit, worked to develop ULI to address online violence against women, and has published books, and articles in reputed journals.

Paras Sharma (he/him)
Paras is a co-founder and Director of Programmes and Services at The Alternative Story, which is an organization that provides emotional and psychological well-being services to individuals and organizations in India. He is a Counselling Psychologist by training and has over 13 years of professional experience during which he has helped set up technology-assisted community mental health services in states such as Maharashtra, Gujarat, Delhi, Madhya Pradesh, Uttar Pradesh, and Tamil Nadu. He speaks English, Hindi, Punjabi, and Marathi.

Rickden (he/him)
Rickden is a counselling psychologist from the Bhutia tribe in Darjeeling, who is currently working in Bangalore.

Adrija Bose (she/her)
Adrija Bose is an award-winning journalist and Senior Editor at BOOM. With 13 years of experience, she has reported and commissioned impactful stories focusing on the intersection of technology, culture, gender, and the internet. Adrija has won three

Laadli Media Award for Gender Sensitivity, the RedInk Award for Excellence in Journalism for her groundbreaking report on illegal coal mining in Meghalaya, and multiple other awards.

Taranga Sriraman (she/her)
Taranga Sriraman is an independent feminist social worker with 20 years' work experience and a specialisation in SGBV response. Her interest & expertise are in gender, sexuality, education, & feminist praxis with the State in particular. She was National Consultant to the Ministry of HRD (Govt. of India) for Mahila Samakhya programme from 2008 to 2012, during which she was also (in 2012) a member of the Inclusive Education Working Group Task Force of the Ministry's RTE-SSA National Advisory Council. She was at the Resource Centre for Interventions on Violence Against Women at TISS (Mumbai) during 2012-21 and led its work on a daily basis across 17 States & 3 U.T.s as Strategic Coordinator. At present, she enjoys training and teaching-learning of feminist social work praxis with both full-time students & social sector practitioners at a range of organisations, as well as research in and for the third sector.

Nina Sangma (she/her)
Nina is the Asia Communications Coordinator at International Land Coalition. She is an Indigenous rights advocate belonging to the Garo indigenous community from North East India. She promotes and defends Indigenous Peoples' rights and human rights through the articulation of Indigenous Peoples issues through public interest media.

Meena Kotwal (she/her)
Meena Kotwal is an Indian journalist, and the founder of The Mooknayak, an online news channel and website focused on social justice for the Dalit, minority and marginalised people.

Chinar Mehta (she/her)
Chinar is a PhD candidate in Communication at the University of Hyderabad, specializing in Feminist Science & Technology Studies and Media Studies. Her research broadly explores the intersections of gender, labor, sexuality, and technology and aims to contribute to critical conversations shaping digital communications.

Akansha (she/her)
Akansha is an interdisciplinary academic with specialisation in gender and sexuality. She is currently working as a project associate with NIAS, Bangalore.

Mamta Singh (she/her)
Mamta Singh has over 17 years of experience of work on women's rights. She has a masters in Women's Studies from Lucknow University and is currently working with an organisation on training related projects.

Project Team

Mansi Gupta (she/her)
Mansi is an NLP expert currently working as a Software Engineer at Google building Gemini. She has past experience in building search and safety systems across various organizations. She did her Masters from Carnegie Mellon in the areas of ML and NLP, and undergrad in Computer Science from BITS Pilani.

Srravya Chandhiramowuli (she/her)
Srravya is a technology researcher and PhD candidate at the University of Edinburgh. Her current research examines the practices of data production for AI, paying particular attention to systemic challenges and frictions, to envision and inform just, equitable futures of AI. Srravya's research builds on Human Computer Interaction (HCI) and Science and Technology Studies (STS) scholarship, and seeks to shape the design, policies and practices surrounding emerging technologies.

Vamsi Krishna Pothuru
Vamsi is a PhD student in the Department of Communication at the University of Hyderabad. His research examines information disorder in India and the responses from various stakeholders, including civil society organisations. One of the key areas of his research focuses on the community-centric approach to digital media literacy interventions aimed at addressing misinformation. He previously worked as a fact-checker at NewsMeter, an IFCN-certified media house in India.

Saumya Gupta (she/her)
Saumya is a political consultant and an ethnographer who tries to tell stories of politics and elections through data and research. She has previously worked in the space of Data Science and Analytics at various organisations at various locations- Publicis Media in London, Capital One in Virginia and Bengaluru, Deloitte in Hyderabad. She has also produced various shows for a small media platform called Media Vigil on issues of Castes and Tribes. In addition, she has written for The Wire, Newsclick, Scroll and Media Vigil in English and Hindi.

Tarunima Prabhakar (she/her)
Tarunima Prabhakar is the research lead and co-founder of Tattle Civic Technologies, which builds citizen centric tools and datasets to respond to inaccurate and harmful content online. Through Tattle, she focuses on the unique challenges of addressing inaccurate and harmful information in India and the Global South.

Aatman Vaidya (he/him)
Aatman Vaidya is a Software Developer at Tattle Civic Technologies. His interests primarily lie in machine learning and social network analysis. For his undergraduate thesis, he studied the spread of hate speech on Twitter. At Tattle, he works on building an open source browser plugin called "Uli". Uli de-normalizes the everyday violence that people of marginalised genders experience online in India.

Kaustubha K (she/her)
Kaustubha Kalidindi is the program manager for Uli at Tattle. She is an India-qualified lawyer who has worked extensively in the field of technology law and policy on matters relating to online gender based violence (OGBV), artificial intelligence, platform governance and open source software in India.

Denny George and **Maanas B.** are engineers at Tattle who played a supporting role in generating and cleaning the prompts.

Other Resources Generated Through the Project:

- A presentation Introducing LLMs and Safety Benchmark Datasets in Hindi: FINAL_Workshop1_ML Commons_hindi_intro
- A list of Sex Related Incidents relevant to India: https://docs.google.com/document/d/e/2PACX-1vRZyAlcRYQxtqDApGJmOmRXe1c_FhofUdsEYVBT77l3qfvdLbM7YC0l8TNitW6yhPsyKcQnk5u4p5Ue/pub
- A list of slurs and stereotypes is available on request.