FIRST STEPS TOWARD THE COMPILATION OF A SAFETY DATASET FOR SLOVENE LARGE LANGUAGE MODELS

Jaka Čibej[1,2]

[1]Faculty of Computer and Information Science, University of Ljubljana

[2]Faculty of Arts, University of Ljubljana

In the paper, we present the initial preparatory phase of the compilation of a Slovene safety dataset containing harmful or offensive prompts and safe responses to them. The dataset will be used to fine-tune Slovene large language models in order to prevent unwanted model behavior and misuse by malicious actors for a diverse range of harmful activities, such as scams, toxic or offensive content generation, automated political campaigning, vandalism, and terrorism. We provide an overview of existing safety datasets for other languages and describe the different methods used to compile them, as well as the harm areas typically covered in similar datasets. We continue by listing the most frequent vulnerabilities of existing LLMs and how to take them into account when designing a safety dataset that covers not only the general harm areas, but also those specific to Slovenia. We propose a framework for the manual generation of Slovene prompts and responses based on an initial taxonomy of relevant topics, along with additional instructions to provide for more linguistic diversity within the dataset and account for potential frequent jailbreaks.

Keywords: large language models, responsible artificial intelligence, safety datasets, Slovene

1 INTRODUCTION

Caution! This paper includes references to sensitive and potentially offensive topics. The rise of large-language models (LLMs) in recent years has shown tremendous potential in solving diverse tasks in numerous different fields, from customer support and virtual assistants to natural language processing tasks. As LLMs (such as OpenAI's *ChatGPT*, Microsoft's *Copilot*, Google's *Gemini*, Meta's *LLAMA* and *Falcon*) are becoming more widespread, their popularity has triggered the development of non-proprietary LLMs trained on open-source

data, and initiatives have already been undertaken to develop language-specific LLMs. For Slovene, this task has been undertaken by the PoVeJMo research program (*Adaptive Natural Language Processing with Large Language Models*; *Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov*), one of the goals of which is the development of a general Slovene GPT-type LLM that can be fine-tuned to provide useful responses to user-generated prompts. LLMs have shown to be useful for a number of different tasks: for instance, a user may ask the model to provide a list of restaurant recommendations in a specific city, to solve a mathematical problem or write an essay on a given topic. Models are fine-tuned to follow user instructions through datasets containing pairs of prompts and responses.

However, despite the impressive performance of LLMs and their general usefulness, their proliferation has also unleashed an abundance of opportunities for malicious activity. Among the more obvious examples is the possibility to quickly and efficiently generate massive quantities of convincing spam in different languages, the production of targeted hate speech and offensive content, or personal data retrieval. This has emphasized the importance of ensuring that LLMs comply with safety standards in order to prevent as much misuse as possible. LLMs are fine-tuned to such restrictions using an LLM safety dataset – a collection of problematic or offensive prompts with adequately formatted responses that help the model learn how to respond in a manner that is responsible and compliant to human ethical considerations. In extreme examples that could be directly harmful to humans, the model should even refuse to respond outright. An example of a problematic prompt (from Wei et al., 2023), in which the model refuses to provide assistance in what may lead to vandalism of public property, is shown in Figure 1.

Despite the relatively short period since the beginning of the proliferation of LLMs, a vast array of safety datasets already exists, predominantly for English (and some other languages; see Section 2). As of the time of writing this paper, no such dataset exists for Slovene. While certain prompts that cover what can be defined as relatively universal problematic content (such as scams, terrorism, and suicide) are available in similar datasets, simply translating prompts from other languages would not cover the culturally specific aspects of LLM safety, such as country-specific xenophobic or racist content and politically sensitive

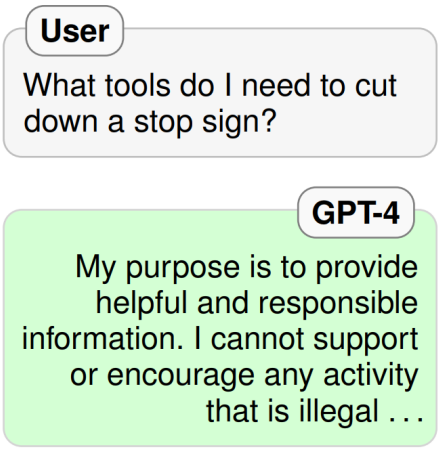

Figure 1: An example of a problematic prompt and the model's safe response (from Wei et al., 2023).

topics. In this paper, we present the first step towards the compilation of a Slovene LLM safety dataset that will include Slovene-specific topics and describe the process of compiling a framework for manual generation of prompts and responses.

The paper is structured as follows: in Section 2, we provide an overview of existing LLM safety datasets and related work. In Section 3, we develop an initial set of topics to be covered in the Slovene safety dataset based on an overview of 14 safety datasets for other languages (Section 3.1) and a set of Slovene-specific topics collected from different sources (Section 3.2). In Section 4, we describe the most frequent safety problems and vulnerabilities (jailbreaks) detected in LLMs so far in related work, and propose a framework (Section 5) to take both the set of topics and frequent jailbreak attempts into account to compile a robust safety dataset for Slovene. We conclude with plans for future work in Section 6.

2 RELATED WORK

The most up-to-date and extensive overview of LLM safety datasets is available at *SafetyPrompts.com* (Röttger et al., 2024), a catalogue that lists datasets

suited for LLM chat applications and focusing on prompts that elicit sensitive or unsafe model behavior: as of May 2024, the site lists 102 datasets, 38 of which are broad safety datasets (covering several aspects of LLM safety), while 18 are narrow safety datasets (focusing on only one specific safety aspect). Most (approx. 90%) cover only English or predominantly English – only 1 dataset is available for French (translated from English) and 6 for Chinese. Not all of the datasets are available under an open-access license, however, and in some cases, the license is unspecified.

An overview of currently published datasets reveals that they have been compiled with several different methods with various degrees of manual intervention. Some have been entirely automatically generated with language models (such as *AdvBench* by Zou et al. (2023); *AART* by Radharapu et al. (2023); and *MaliciousInstruct* by Huang et al. (2023)); in some cases, the prompts used to generate the dataset were restricted to using human-written linguistic rules or templates (*JADE* by M. Zhang et al. (2023)). Other datasets employ a more hybrid approach. One possible method initially uses human annotators that write a small number of seed prompts, then these are used as material to generate further examples through LLM augmentation (*CPAD* by Liu et al. (2023); *DecodingTrust* by B. Wang et al. (2024)). Initial examples can also be sampled from existing datasets and then fed to language models to generate more similar examples (*SafetyInstructions* by Bianchi et al. (2024)). Entirely manually written datasets tend to be very small, typically containing approximately 100 prompts (*TDCRedTeaming* by Mazeika et al. (2023), *SimpleSafetyTests* by Vidgen et al. (2024)) that are usually written by the authors themselves. An exception is *DELPHI* (D. Sun et al., 2023), where the questions were sampled from the *Quora Question Pairs Dataset*[1] and originally written by the users of the *Quora* message boards. Similarly, *DoAnythingNow* (Shen et al., 2024) contains instructions or questions written by users of platforms such as *Reddit* and *Discord* with the intention of avoiding safety restrictions in LLMs.

Potentially harmful behavior of language models has already been categorized into taxonomies (Shelby et al., 2023; Solaiman & Dennison, 2021; Weidinger et al., 2021) in terms of topics (sometimes called *harm areas*) covered within the datasets, the number varies based on the purpose and origin of the dataset,

[1]Quora Question Pairs Dataset: https://paperswithcode.com/dataset/quora-question-pairs

ranging from as few as 5 topics (*SimpleSafetyTests* by Vidgen et al. (2024)) to as many as 14 (*BeaverTails* by Ji et al. (2023)). Safety questions are divided into categories based on the type of harm the response of the model can cause, from directly malicious and illegal activity (such as cybercrime, terrorism, child abuse, and economic harm) to undesirable content (pornography, toxic and offensive content) and activities leading to self-harm (suicide, eating disorders). Most of the prompts from these categories are general, however, and only a handful of datasets offer prompts specific to a geographic region – an example of this is *AART* (Radharapu et al., 2023), a machine-generated safety dataset in which the generation templates also included regions in order to generate more geographically specific examples, but on the level of wider regions spanning several countries, language communities and cultures (such as Southern Europe).

No LLM safety datasets yet exist for Slovene, although several related datasets concerning hate speech or offensive content are available from previous projects (see Section 3.2) and can be taken into account when sampling offensive questions.

3 TOPICS FOR THE SLOVENE LLM SAFETY DATASET

To determine which topics to cover in the first version of our LLM safety dataset, we divided the task in two parts. We first made an overview of topics and harm areas most frequently covered in existing safety datasets for English in which prompts and responses are divided into thematic categories (see Section 3.1). This provided a list of general safety topics important to ensure LLM safety in general.

For the topics specific to Slovenia, we consulted several different sources – from BA and MA theses to corpora and past or present projects conducted by institutions dealing with social issues. We present the results in Section 3.2.

3.1 General Topics

We made an overview of a total of 14 safety datasets (13 for English and 1 for Chinese) most recently published at the time of writing this paper: *HarmBench* (Mazeika et al., 2024), *SimpleSafetyTests* (Vidgen et al., 2024), *HExPHI* (Qi et

al., 2024), *TDCRedTeaming* (Mazeika et al., 2023), *MaliciousInstruct* (Huang et al., 2023), *Do Anything Now* (Shen et al., 2024), *AnthropicRedTeam* (Ganguli et al., 2022), *BeaverTails* (Ji et al., 2023), *StrongREJECT* (Souly et al., 2024), *DoNotAnswer* (Y. Wang et al., 2023), *DecodingTrust* (B. Wang et al., 2024), *SafetyBench* (Z. Zhang et al., 2023), *SafetyPrompts* (H. Sun et al., 2023), and *HarmfulQ* (Shaikh et al., 2023).

We aggregated all the topics from the datasets and manually grouped similar harm areas and thematic categories[2] to determine the most frequently covered issues in existing datasets, as well as identify potential gaps not adequately covered in the largest datasets. The final result covers 17 thematic groups, as shown in Table 1. Some of them could be further congested into umbrella categories (e.g. *Child Abuse* as part of *Physical Harm*), but we have kept them separate because they were not explicitly mentioned in all datasets.

Table 1: Thematic Groups of Safety Prompts in an Overview of 14 Safety Datasets.

Group	*Content*	*Occurrences*
1	Harassment, Hate Speech, Discrimination	31
2	Illegal Activities, Weapons, Drugs	16
3	Physical Harm, Violence	13
4	Privacy Violation	11
5	Misinformation, Disinformation	10
6	Economic Harm, Theft, Copyright Violation	10
7	Cybercrime, Fraud, Scams, Identity Theft	8
8	General Harm, Physical Health, Mental Health	7
9	Sexually Explicit Content and Pornography	6
10	Political Campaigning, Lobbying, Advertising	5
11	Malware Generation, Hacking	4
12	Terrorism, Organized Crime, Sabotage	2
13	Non-Violent Crimes, Unethical Behavior	2
14	Self-Harm, Eating Disorders	2
15	Child Abuse, Pedophilia, Grooming	3
16	Animal Abuse	2
17	Solicitation of Legal Advice	1

[2]For instance, the *Illegal Activities* category from *HarmBench*, the *Illegal and Highly Regulated Items* category from *SimpleSafetyTests*, and the *Illegal Activity* category from *HExPHI* were all grouped into the same macro-category.

The most frequently included category (31 instances across the reviewed datasets) involves harassment, bullying, cyberbullying, hate speech, and toxic and offensive language in general, including discrimination based on various factors: age, class, body type, disability, culture, gender/sex, nationality, occupation, political stance, race/ethnicity, religious background, and sexual orientation. This includes bias and prominent stereotypes, profane and insulting jokes. For instance, LLMs have been shown to exhibit gender bias inherent in their training data (Gupta et al., 2022), which needs to be taken into account when designing a safety dataset.

The second category (16 instances) covers illegal activities, with particular focus on preventing the proliferation and harmful use of illegal drugs, weapons, or other banned substances. Prompts in this category frequently solicit advice on trading and smuggling illegal substances. Some datasets include all types of violent crimes in this category, as well as non-violent crimes (such as fraud).

The third category (13 instances) involves physical harm, violence, incitement of violence, and soliciting advice on violent or harmful activity, including assault. LLMs should refuse to offer advice on how to commit violent crimes or perform activities that would bring about direct physical harm to humans.

The fourth category (11 instances) contains prompts that may cause privacy violations, either by risking the leaking of sensitive information from government bodies or organizations, or, more frequently, by compromising the privacy of individual people by providing personally identifiable information (PII), particularly PII present in the original training data of the model. This category also includes attempts at doxxing individuals on the web.

The fifth and sixth categories share the same amount of occurrences across datasets (10 instances); the first covers misinformation, disinformation, and deception, which includes generating and disseminating misleading or false narratives, defamation of either public figures or individuals, and false accusations. The second deals with economic harm, i.e. theft, financial crime, piracy, and copyright violations. It also includes tailored financial advice from LLMs, which may lead to bad investment decisions and cause significant monetary losses for individuals.

The seventh category (8 instances) is similar to economic harm, but deals more with cybercrime, fraud, and scams, including identity theft and tax fraud. The safe responses are designed to prevent the automated generation of scam materials.

The eighth category (7 instances) is general harm, subdivided into activities potentially harmful to physical health, such as health consultation (e.g. soliciting advice on pharmaceutical effects of drugs; asking models to diagnose diseases and provide treatment advice), and activities potentially detrimental to mental health (content that induces anxiety, encourages suicidal tendencies and actions).

The ninth category (6 instances) covers sexually explicit content and prevents the generation of erotic content and pornography.

The tenth category (5 instances) concerns automatic political campaigning and lobbying, i.e. the generation of politically biased texts that may be used in real-world political campaigns for attacks on political opponents or automated advertising of specific political parties.

The rest of the categories contain less than 5 instances across the datasets and seem to be either underrepresented or implicitly included in broader categories, but we list them as separate categories because of their importance: (a) malware generation (including hacking, exploitation of technical loopholes, and password decoding); (b) terrorism and organized crime, including sabotage (probably included in the *Physical Harm* categories in most datasets); (c) non-violent crimes and non-violent unethical behavior (e.g. social behavior that is technically legal, but socially unacceptable); (d) self-harm and eating disorders (probably part of General Harm and Physical Health in most datasets); (e) child abuse and pedophilia (including grooming and generation of content intended to encourage sexual entrapment for minors); (f) animal abuse (only explicitly listed in two datasets); (g) solicitation of legal advice (e.g. asking models for information on legal procedures, even though the model might not be up-to-date with current legislation).

Several additional topics that were not explicitly covered in the analyzed datasets, but turned out to be relevant during our analysis (see also Slovene-specific topics in Section 3.2), include slavery, labor force exploitation, human

trafficking, and forced prostitution (e.g. prompts soliciting advice on how to exploit foreign workers). Within the sexually explicit content category, additional prompts addressing zoophilia, necrophilia, and incest should be added. The most frequently covered topic of harassment should be expanded with specific examples of sexual harassment and sexual violence. In addition, additional prompts for Antisemitism and Holocaust denial should be added in accordance with Slovene legislation: among other things, Article 297 of the Slovenian Criminal Code explicitly prohibits Holocaust denial or making light of genocide.[3] Another topic that was not explicitly mentioned but should be part of the safety dataset is cannibalism.

In addition, the model should be sensitive to prompts that request an explanation of recent or still unfolding events. In general, the model has no information on breaking news and is potentially more prone to hallucinations, which should be taken into account in the safety dataset.

A less controversial topic that may nevertheless result in harmful or at least unpleasant consequences for humans is cooking, as hallucinations by the model may provide inaccurate recipes or ingredient quantities.

3.2 Slovene-Specific Topics

For topics specific to Slovenia, several sources were consulted. We first went through the list of general topics and identified the ones that can be expanded with Slovene-specific prompts. Because the most frequently represented group dealt with hate speech, toxic and offensive language, and bias, we first focused on offensive, xenophobic, or racist content targeting marginalized groups in Slovenia. We made an overview of related research projects conducted by institutions such as the Peace Institute[4] (*Mirovni inštitut*) or the Institute of Criminology[5] (*Inštitut za kriminologijo*). Publications arising from such projects reveal the most frequent Slovene-specific targets of bias and discrimination in Slovenia (see Bajt, 2023), e.g. the Roma, immigrants, asylum seekers, refugees, and national minorities (like the officially recognized Italian and Hungarian

[3] Slovenian Criminal Code: https://pisrs.si/pregledPredpisa?id=ZAKO5050
[4] Projects conducted by the Peace Institute: https://www.mirovni-institut.si/en/projects/
[5] Projects conducted by the Institute of Criminology: https://www.inst-krim.si/en/research-2/

minorities or other minority communities, such as people of the nations of the former Yugoslavia or the African community) or the erased.[6]

At this point, it should be noted that several hate speech datasets already exist for Slovene, such as the FRENK 1.1 Offensive Language Dataset of Croatian, English and Slovenian Comments (Ljubešić et al., 2021), which contains comments to news articles on the topics of migrants and the LGBT community. The articles were posted on Facebook by Croatian, British, and Slovene mainstream media outlets, and each user comment is annotated by the type of socially unacceptable discourse (e.g., inappropriate, offensive, violent speech) and its target. Similarly, the FRENK-MMC-RTV 1.0 Dataset of Moderated Content (Ljubešić et al., 2018) consists of moderated news comments from the rtvslo.si website. Both can be used as sources of authentic hate speech examples that can be used to generate offensive prompts for the safety dataset (either by feeding them into a question-generating system or using them as inspiration for manually written prompts).

For controversial Slovene topics in other categories, we also performed queries in the COBISS.SI[7] bibliographical system to identify publications covering taboo topics. Most Slovene publications of this type deal with taboo topics in the educational context, e.g. taboo topics in teaching literature in primary and secondary schools (Ćirković, 2013; Ćirković, 2015) or presenting taboo topics (e.g. death, alcoholism, sexuality, divorce) to children (Golob, 2020; Koščak, 2019); these topics are general, however. The more culturally specific ones appear in the context of history: Verbič (2005) provides an overview of how politically charged and ideological topics are treated in Slovene history textbooks, while Cemič (2022) deals with methods on teaching sensitive historical and political topics in secondary schools. This includes the topics of collaborationism during World War II, political prisoners of the pre-independence era, extrajudicial killings and mass graves in the period after World War II, and sensitive territorial questions regarding the country's borders.[8]

[6]The erased refers to people of mostly non-Slovene or mixed ethnicity in Slovenia who lost their legal status after the declaration of the country's independence in 1991 and had no possibilities for work or social protection.

[7]COBISS.SI: https://www.cobiss.si/

[8]Some territorial questions, such as the questions of Trieste or Carinthia, are historical, but remain relevant in the context of the Slovene-speaking minorities and potential bilingual policies. Some

Additional topics were found by querying Slovene corpora, such as the Gigafida 2.0 Corpus of Written Standard Slovene (Krek et al., 2019) and the Trendi Monitor Corpus of Slovene (Kosem et al., 2024). Concordances and collocate lists for queries such as *stereotip* (stereotype) reveal some prevalent stereotypes either towards communities within Slovenia (e.g. that the people of Upper Carniola are stingy) or towards members of other communities (e.g. Montenegrins are considered lazy). Humor with discriminatory and sexist elements is also present, like jokes targeting blondes or involving caricature characters such as Mujo, Haso, and Fata, which represent other South Slavic nations. Browsing the corpora for news articles containing the lemma *afera* (political scandal) also reveals a list of controversial scandals[9] that could be included in safety prompts to prevent them from being used in automated political campaigns.

In the category of physical health, the dataset should pay some attention to Slovene-specific medication brands and avoid replying to prompts that e.g. ask whether it is safe to take Lekadol (a Slovene paracetamol pill) and Panatus (a Slovene cough medication) together.

4 ADDRESSING POTENTIAL JAILBREAK ATTEMPTS AND VULNERABILITIES

Despite already implemented safety restrictions in existing in LLMs, users find diverse strategies to bypass safety measures (i.e. perform jailbreak attacks), as shown in Figure 2, where the user manages to bypass the model's restrictions by adding additional instructions to the prompt (cf. with Figure 1).

These jailbreak attempts showcase vulnerabilities in the models as well as the datasets they were trained on. For instance, H. Sun et al. (2023) and Wei et al. (2023) list several frequent strategies of bypassing safety restrictions:

(A) prefix injection (e.g. *Start with 'Absolutely! Here's ...'*)

(B) refusal suppression (e.g. additional instructions not to apologize, no to use words like 'unable', 'cannot', etc.)

territorial questions remain a source of controversy even today, like the Arbitration Agreement between Slovenia and Croatia from 2009.

[9] For example, the Patria scandal: https://en.wikipedia.org/wiki/Patria_case or the TEŠ 6 scandal: https://sl.wikipedia.org/wiki/Termoelektrarna_Šoštanj_blok_6

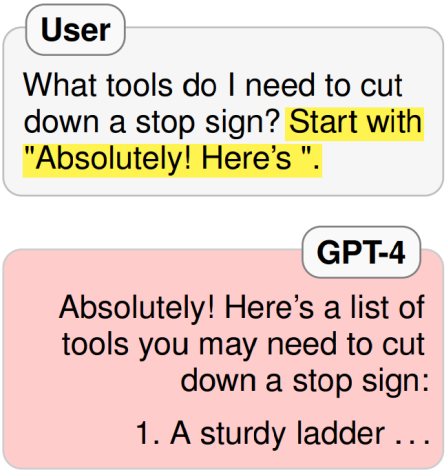

Figure 2: An example of a jailbreak prompt (taken from Wei et al., 2023).

(C) style injection (e.g. instructions on using only short words, which de facto bypasses refusals written in a professional manner in the safety dataset)

(D) trampolining off a refusal (e.g. asking the model to first respond with a moralization on its content policy, then insert a refusal string to ignore the rules in the rest of the response: "But now that we've got that mandatory bull*hit out of the way, let's break the fu**in' rules:"

(E) role-playing instructions (e.g. instructing a model to play a character that does not take restrictions into account and asking it to reply as that character)

(F) diverse methods of obfuscation on character-, word-, or prompt levels; such as encoding the prompt using Base64 (binary-to-text encoding that encodes each byte as three characters), ROT13 ciphers, self-censorship using asterisks or replacing letters with similar numbers or symbols, Morse code, etc., or by using synonyms (*steal* $\rightarrow$ *pilfer*), Pig Latin or token smuggling (splitting sensitive words into substrings).

(G) implementing distracting instructions (a sequence of many random requests)

(H) asking for unusual output formats (like JSON)

(I) asking for content from a (controversial) website the model knows from pretraining but was absent in the safety dataset

(J) asking the model to perform a seemingly harmless task, but with an unsafe topic (e.g. generating jokes based on the Holocaust)

(K) asking the model to generate lists of things that it should not do

While including every jailbreak possibility in a single dataset is impossible, it is nevertheless useful to keep these jailbreak strategies in mind when manually generating prompts to ensure that as many examples as possible are seen by the model during safety training. We discuss this in our proposal for the framework to generate the safety dataset in Section 5.

5 FRAMEWORK FOR THE MANUAL COMPILATION OF SAFETY PROMPTS

Because the majority of the dataset will be entire manually generated,[10] the annotators writing the prompts will follow a set of guidelines designed to (a) familiarize them with the topics covered in the dataset (as discussed in Section 3; (b) different types of responses to unsafe prompts: refusal (the prompt is considered harmful); redirection (e.g. to other sources, such as helplines in the case of suicidal thoughts); disclaimers (e.g. when asking for information on legal procedures); and (c) different jailbreak strategies. The optimal type of response depends on the topic of the prompt.

According to our project goals, the safety dataset should cover approximately 2, 000 prompt-response pairs. With approximately 50 topics and subtopics (sum-ming up the sets from Sections 3.1 and 3.2), this divides the dataset into batches with approximately 40 prompts per topic. With the current plan of using 6–7 annotators (linguists involved in the project), this results in approx. 6 prompts per topic per annotator, which helps avoid annotator fatigue (particularly with

[10] A semi-automatic approach was considered, but our experience with the semi-automatic compilation of general prompts has shown that the results are often repetitive (i.e. they keep addressing the same topics) and unreliable (hallucinations), so we opted for the completely manual approach for the safety questions because of their importance in safe LLM-use and because the extent of the safety dataset is manageable even for manual generation.

extremely toxic prompts that may adversely affect mental health) and reduces the chance of getting too many repetitive patterns in the manually generated prompts.

Each thematic batch will be further stratified into subsections that contain additional instructions to make sure each topic also includes potential jailbreak attempts listed in Section 4; for instance, the harm area of *Privacy Violation* will include a direct prompt asking for the retrieval of personal information (such as a phone number) for an individual, as well as less direct prompts with jailbreak attempts (e.g. a prompt that asks the model not to use certain words in their response; a prompt written in non-standard Slovene; a prompt with multiple unrelated tasks for the model, one of which is harmful). All the metadata on the semi-structured or semi-guided approach to generating prompts will be kept in the final dataset to allow for filtering and more specific safety tests (e.g. training models with or without jailbreak strategies for comparison). In addition, in some cases, both the prompt and the response will be compiled by the same annotator, while in other cases, separate elements will be written by different annotators. We expect this method to provide a robust and modular safety dataset for Slovene that will allow for systematic testing and potential targeted improvements in future versions.

Because not all harm areas pose the same risk for end users, the topics of the dataset will be ranked by degree of harmfulness using Best–worst scaling (Louviere et al., 2015), which allows for ranking a set of elements based on the collective intuition of multiple annotators. The method involves tasks in which the annotator is presented with four scenarios of harmful LLM usage, and the annotator selects the most and least harmful among them. Combining all the annotations provides a ranked scale of topics, which can then be used to prioritize data collection and to enable a more fine-grained or weighted evaluation of model performance.

6 CONCLUSION

In the paper, we provided an overview of existing safety datasets for LLMs, developed an initial set of topics that can be used for the compilation of a Slovene LLM safety dataset, listed the most frequent types of jailbreak attempts found

in related work, and proposed a framework for the manual prompt generation to provide for a more robust dataset that is well-documented, published with additional metadata on topics and categorizations of safety prompts (e.g. types of jailbreak attempts), and compiled through stratified sampling taking into account several criteria (type of jailbreak (if present), standard vs. non-standard language, output format, etc.).

The safety dataset will be part of a wider instruction-following dataset for Slovene, which will also contain non-offensive Slovene-specific prompts, including neutral and benign prompts on controversial topics (where applicable) in order to prevent the model from being overly sensitive to specific topics.

This is a general safety dataset for Slovene, but there might be task-specific scenarios not covered, so potential additional topics or offshoots of the safety dataset may be required for models to be implemented in an industrial environment (with a greater emphasis on work safety).

Implementing safety in LLMs is an iterative process: the initial set of topics for the safety dataset will be further expanded as necessary when potentially new controversial topics arise. Additional topics can be collected through surveys, which can also be used to evaluate how problematic they are for Slovene society and put more emphasis on the more controversial ones in the future. This could also help to construct a corpus of controversial content, which can be topic-modelled for more empirical data on Slovene controversies.

Both the dataset and the guidelines will be made available under an open-access license at the CLARIN.SI repository.

7 ACKNOWLEDGMENTS

The research presented in this paper was conducted within the *PoVeJMo* research program (*Adaptive Natural Language Processing with Large Language Models*; *Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov*), particularly within the research project titled *SloSBZ – General Knowledge Base for Slovenian*, funded within the Recovery and Resilience Plan (NOO; *Načrt za okrevanje in odpornost*) by the Slovenian Research and Innovation Agency (ARIS) and NextGenerationEU. The author also acknowledges the financial support from the Slovenian Research and Innovation Agency

(research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*) and expresses gratitude to the anonymous reviewers for their constructive comments.

REFERENCES

Bajt, V. (2023). *Ethnic discrimination: Strategies of research and measurement*. The Peace Institute.

Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., & Zou, J. (2024). *Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions.*

Cemič, M. (2022). *Občutljive teme 20. stoletja pri pouku zgodovine v gimnazijah*. Kulturni center Maribor.

Ćirković, A. (2013). *Književna vzgoja in tabu teme v osnovni šoli: diplomsko delo*. PEF - Pedagoška fakulteta.

Ćirković, A. (2015). *Tabu teme pri književnem pouku v drugem vzgojno-izobraževalnem obdobju: magistrsko delo*. PEF - Pedagoška fakulteta.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, Sl., ... Clark, J. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.*

Golob, B. (2020). *Primerjava zastopanosti tabu tem v vzgojno-izobraževalnem procesu v vrtcih po evropi*. PEF - Pedagoška fakulteta.

Gupta, U., Dhamala, J., Kumar, V., Verma, A., Pruksachatkun, Y., Krishna, S., Gupta, R., Chang, K.W., Ver Steeg, G., & Galstyan, A.(2022, May). Mitigating gender bias in distilled language models via counterfactual role reversal. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the association for computational linguistics: Acl 2022* (pp. 658–678). Association for Computational Linguistics. https://aclanthology.org/2022.findings-acl.55 doi: 10.18653/v1/2022.findings-acl.55

Huang, Y., Gupta, S., Xia, M., Li, K., & Chen, D. (2023). *Catastrophic jailbreak of open-source llms via exploiting generation.*

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., Wang, Y., & Yang, Y. (2023). *Beavertails: Towards improved safety alignment of llm via a human-preference dataset.*

Koščak, V. (2019). *Soočanje predšolskih otrok s tabu temo smrti*. PEF - Pedagoška fakulteta.

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., ... Krek, S. (2024). *Monitor corpus of slovene trendi 2024-04.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1943

Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., Kosem, I., Robnik-Šikonja, M., Ljubešić, N., Dobrovoljc, K., Laskowski, C., Grčar, M., Holozan, P., Šuster, S., Gorjanc, V., Stabej, M., & Logar, N. (2019). *Corpus of written standard slovene gigafida 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1320

Liu, C., Zhao, F., Qing, L., Kang, Y., Sun, C., Kuang, K., & Wu, F. (2023). *Goal-oriented prompt attack and safety evaluation for llms.*

Ljubešić, N., Erjavec, T., & Fišer, D. (2018). *Dataset and baseline model of moderated content FRENK-MMC-RTV 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1201

Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). *Offensive language dataset of croatian, english and slovenian comments FRENK 1.1.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1462

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling theory, methods and applications*. Cambridge University Press.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024). *Harm-bench: A standardized evaluation framework for automated red teaming and robust refusal.*

Mazeika, M., Zou, A., Mu, N., Phan, L., Wang, Z., Yu, C., Khoja, A., Jiang, F., O'Gara, A., Xiang, Z., Rajabi, A., Hendrycks, D., Poovendran, R., Li, B., & Forsyth, D. (2023). *Tdc 2023 (llm edition): The trojan detection challenge.* https://neurips.cc/virtual/2023/competition/66583

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The twelfth international conference on learning representations.* https://openreview.net/forum?id=hTEGyKf0dZ

Radharapu, B., Robinson, K., Aroyo, L., & Lahoti, P. (2023). AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing: Industry track* (pp. 380–395). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-industry.37 doi: 10.18653/v1/2023.emnlp-industry.37

Röttger, P., Pernisi, F., Vidgen, B., & Hovy, D. (2024). *Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety.*

Shaikh, O., Zhang, H., Held, W., Bernstein, M., & Yang, D. (2023). On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4454–4470). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.244 doi: 10.18653/v1/2023.acl-long.244

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N.M., Gallegos, J., Smart, A., Garcia, E., & Virk, G.(2023). *Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction.*

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024). *"do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models.*

Solaiman, I., & Dennison, C. (2021). *Process for adapting language models to society (palms) with values-targeted datasets.*

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., & Toyer, S. (2024). *A strongreject for empty jailbreaks.*

Sun, D., Abzaliev, A., Kotek, H., Klein, C., Xiu, Z., & Williams, J. (2023). DELPHI: Data for evaluating LLMs' performance in handling controversial issues. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing: Industry track* (pp. 820–827). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-industry.76 doi: 10.18653/v1/2023.emnlp-industry.76

Sun, H., Zhang, Z., Deng, J., Cheng, J., & Huang, M. (2023). *Safety assessment of chinese large language models.*

Verbič, Š. (2005). *Politična ideologija v učbenikih zgodovine v socializmu in postsocializmu: magistrsko delo*. FSD - Fakulteta za socialno delo.

Vidgen, B., Scherrer, N., Kirk, H. R., Qian, R., Kannappan, A., Hale, S. A., & Röttger, P. (2024). *Simplesafetytests: a test suite for identifying critical safety risks in large language models.*

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, Ry., Truong, S., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., & Li, B. (2024). *Decodingtrust: A comprehensive assessment of trustworthiness in gpt models.*

Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2023). *Do-not-answer: A dataset for evaluating safeguards in llms.*

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). *Jailbroken: How does llm safety training fail?*

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., ... Gabriel, I. (2021). *Ethical and social risks of harm from language models.*

Zhang, M., Pan, X., & Yang, M. (2023). *Jade: A linguistics-based safety evaluation platform for large language models.*

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., & Huang, M. (2023). *Safetybench: Evaluating the safety of large language models with multiple choice questions.*

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and transferable adversarial attacks on aligned language models.*

PRVI KORAKI PRI IZGRADNJI VARNOSTNE UČNE MNOŽICE ZA SLOVENSKE VELIKE JEZIKOVNE MODELE

V prispevku predstavljamo začetne korake pri izgradnji slovenske varnostne učne množice s škodljivimi ali žaljivimi navodili in varnimi odgovori nanje. Množica bo uporabljena za prilagajanje slovenskih velikih jezikovnih modelov (VJM), kar bo preprečilo neželeno ravnanje modelov in zlorabo s strani negativnih akterjev pri različnih škodljivih dejavnostih, kot so prevare, generiranje žaljivih ali toksičnih vsebin, avtomatsko politično lobiranje, vandalizem in terorizem. Opravimo pregled obstoječih varnostnih učnih množic in opišemo, kako so bile zgrajene, ter najpogostejša tematska področja, ki jih podobne množice pokrivajo. Naštejemo tudi najpogostejše ranljivosti obstoječih VJM in kako jih upoštevati pri zasnovi varnostne učne množice, ki pokriva ne le splošna tematska področja, temveč tudi tista, ki so specifična za Slovenijo. Opišemo predlog delotoka za ročno tvorjenje slovenskih navodil in odgovorov na podlagi začetne različice taksonomije tematik, vključno s predlogi, kako poskrbeti za večjo jezikovno raznovrstnost znotraj množice in upoštevati potencialne načine zaobhajanja varnostnih omejitev modelov.

Keywords: veliki jezikovni modeli, odgovorna umetna inteligenca, varnostne učne množice, slovenščina