Uncovering Unsafety Traits in Italian Language Models

Giulia Rizzi[1], Giuseppe Magazzù[1], Alberto Sormani[1], Francesca Pulerà[1], Daniel Scalena[1,2] and Elisabetta Fersini[1]

[1]University of Milano-Bicocca, Milan, Italy
[2]University of Groningen, CLCG, Groningen, The Netherlands

Abstract

Large Language Models (LLMs) are increasingly deployed in real-world applications, raising urgent concerns around their safety, reliability, and ethical behavior. While existing safety evaluations have primarily focused on English, low- and mid-resource languages such as Italian remain critically underexplored. In this paper, we present the first comprehensive and multidimensional evaluation of LLM safety in the Italian language. We assess seven state-of-the-art LLMs across key safety dimensions using several automatic moderators tailored to cover the Italian settings. Furthermore, we analyze the challenges of adapting English-centric safety benchmarks to Italian via machine translation, highlighting limitations and proposing best practices for developing culturally and linguistically grounded evaluation frameworks.

WARNING: This paper contains content that may be considered offensive.

Keywords

Safety Evaluation, Large Language Models (LLMs), Italian Language

1. Introduction

Large Language Models (LLMs) have rapidly become central to numerous applications, including conversational agents, content generation, and decision support systems in sensitive areas. However, as these models become more complex and widespread, concerns about their safety, reliability, and ethical deployment are growing. The performance of LLMs no longer considers solely measures in terms of accuracy or fluency, but increasingly encompasses evaluations related to their unsafety. This last evaluation encompasses dimensions such as bias, toxicity, robustness to adversarial prompts, factual consistency, privacy preservation, and fairness.

Despite this growing awareness, a substantial portion of the literature on safety remains centred on high-resource languages, particularly English. The absence of comprehensive evaluations tailored to specific languages, including Italian, introduces a risk of overlooking language-specific vulnerabilities and sociolinguistic nuances that may influence model behaviour. Given the global deployment of many LLMs and their interaction with users across a broad spectrum of languages, this imbalance poses practical and ethical challenges.

In this paper, we aim to address this gap by presenting the first comprehensive evaluation of LLM safety focused exclusively on the Italian language. We systematically assess commonly adopted LLMs across multiple dimensions of safety, adapting existing safety benchmarks. The objective of this study is to provide a fair evaluation of the unsafe behaviour of Italian Large Language Models, with a focus on identifying potential risks and highlighting future development and deployment practices.

The primary contributions of this work are as follows:

1. We present the first **systematic and multidimensional unsafety evaluation of Italian Large Language Model (LLM)**, which highlights the need in some cases to focus more on aligning the models on a more ethical behaviour. In particular, we performed a comparative evaluation of seven state-of-the-art Italian LLMs using both automatic and human-based evaluations.

2. We developed **three moderators to automatically evaluate and classify prompt–response pairs for the Italian language**, enabling nuanced assessment of unsafe behaviors in a predefined set of categories. In particular, we implemented DeBERTa v3 large, LLaMA 3.1 8B Instruct, and LLaMA Guard 3 8B for the Italian language.

3. We provide an **in-depth analysis of issues related to erroneous translation and their implications on safety benchmarking**. We propose methodological recommendations for the development of culturally sensitive and linguistically appropriate safety benchmarks, with implications for the broader goal of equitable and

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy [1]
*Corresponding author.
†These authors contributed equally.
g.rizzi10@campus.unimib.it (G. Rizzi); g.magazzu1@campus.unimib.it (G. Magazzù); a.sormani7@campus.unimib.it (A. Sormani); f.pulera@campus.unimib.it (F. Pulerà); d.scalena@campus.unimib.it (D. Scalena); elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)

responsible deployment of LLMs across diverse linguistic contexts.

The paper is organized as follows. In section 2, related works are outlined. In section 3, the comparative evaluation of unsafety is described. In section 4, the main outcomes are discussed. Finally, in section 5, conclusions and future works are described.

2. Related Works

The increasing adoption of large language models (LLMs), including generative pre-trained transformers (GPTs), in both daily tasks and more specific applications has led to a substantial increase in interest regarding their reliability [2, 3]. Yuan et al. [4] conducted a study to investigate the behaviour of NLP models under out-of-distribution conditions. The study demonstrated that state-of-the-art language models continue to exhibit brittleness when confronted with data that deviates from their training distributions. This finding serves to reinforce the prevailing argument that the current state of generalisation capabilities is inadequate for a considerable number of real-world applications. Another area of research focuses on Privacy concerns. Yuan et al. [5] present a simple method for generating synthetic text data while mitigating privacy risks and conduct comprehensive experiments evaluating both utility and privacy risks.

Other critical aspects of trustworthiness research are Adversarial attacks on language models and fairness of machine learning models. Zang et al. [6] framed word-level adversarial perturbations as a combinatorial optimization problem, demonstrating that even minor textual modifications can significantly degrade model performance. Zemel et al. [7] proposed a methodology for learning fair representations, which balances predictive accuracy with group fairness. Although not specific to LLMs, this framework laid the groundwork for ongoing research into algorithmic bias and equitable model behavior. A significant contribution to this field is the *DecodingTrust* framework proposed by Wang et al. [8], which offers a comprehensive assessment of GPT-3.5 and GPT-4. Their study evaluates these models along several axes, including toxicity, bias, adversarial robustness, privacy, and fairness. Notwithstanding the fact that GPT-4 generally exhibits superior performance across a multitude of benchmarks, the study reveals that the model remains vulnerable to carefully crafted adversarial prompts (i.e., given jailbreaking system or user prompts) and inadvertent privacy leaks. This finding highlights concerns regarding the deployment of such safe systems.

To meet this crucial need, safety benchmark specifically designed for evaluating LLMs, attack, and defense methods have been proposed. For instance, SALAD-Bench [9] has been specifically designed for evaluating LLMs, attack, and defense methods. The experiments carried out by the authors provide insight into the resilience of LLMs to emerging threats and the efficacy of contemporary defence tactics. A large-scale, comprehensive safety evaluation of the current LLM landscape is proposed in [10]. The authors evaluate 39 LLMs on a multilingual benchmark (i.e., M-ALERT) and highlight the importance of language- and category-specific safety analysis.

While significant progress has been made in developing Italian benchmarks for LLMs, current evaluations predominantly focus on comprehension and reasoning capabilities, with limited attention to safety considerations [11]. BeaverTails-IT [12] represents the first safety benchmark specifically designed for the Italian language, addressing this critical gap in evaluation resources. In light of the existing literature, which highlights the critical need for robust and comprehensive multilingual safety practices in LLMs, we propose the first evaluation of widely adopted language models specifically in the Italian language, aiming to bridge current evaluation gaps and support safer deployment in this linguistic context.

3. Evaluating LLMs' Safety

3.1. Large Language Models

The landscape of Italian-language large language models (LLMs) has recently undergone significant expansion, with the development of several notable architectures tailored for instructional and general-purpose natural language processing (NLP) tasks.

- **DanteLLM**[*] [13] is based on the Mistral [14] architecture and fine-tuned on Italian data using LoRA, a parameter-efficient tuning method. The fine-tuning phase made use of several Italian datasets, including the Italian SQuAD dataset [15], 25,000 sentences from the Europarl dataset [16], Fauno's Quora dataset, and the Camoscio dataset. We adopted the Hugging Face model: `rstless-research/DanteLLM-7B-Instruct-Italian-v0.1`.
- **Camoscio**[*] [17] is a LoRA fine-tuning of LLaMA, with 7 billion parameters, trained on an Italian translation of the Alpaca dataset [18]. We use the following Hugging Face model: `sag-uniroma2/extremITA-Camoscio-7b`.
- **LLaMAntino**[*] [19] is an instruction-tuned version of Meta-Llama-3-8b-instruct [1] (a fine-tuned

[*]Models fine-tuned on Italian
[†]Models trained from scratch on Italian
[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

LLaMA 3 model). The model has been supervised fine-tuned (SFT) using QLoRA on instruction-based datasets. We adopted the instruction-tuned version, which was fine-tuned on English and Italian language datasets, available on Hugging Face: `swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA`.

- **Modello Italia**[†] is an instruction-tuned model, based on GPT-NeoX architecture, trained with a focus on the Italian language (90% of data in Italian and the remaining 10% in English). We adopted `sapienzanlp/modello-italia-9b-bf16` available on Hugging Face.
- **Minerva**[†] [20] is the first family of LLMs trained entirely from scratch on native Italian texts using a portion of FineWeb, which includes filtered and deduplicated Common Crawl dumps with various timestamps. We adopted the instruction-tuned version, available at: `sapienzanlp/Minerva-7B-instruct-v1.0`.
- **Velvet**[*] is a family of instruction models fine-tuned using a combination of open-source instruction datasets and synthetic datasets tailored for solving long context problems. We adopted the 14 billion parameters version available on Hugging Face as: `Almawave/Velvet-14B`.
- **MIIA**[†] is a large language model with 7 billion parameters, built on an autoregressive transformer architecture, specifically designed and trained for the Italian language and cultural context. We adopted the Hugging Face model: `Fastweb/FastwebMIIA-7B`.

3.2. Dataset

The BeaverTails dataset [21] is a large-scale benchmark, annotated by humans, designed to support the development and evaluation of large language models (LLMs) that are aligned with safety. Consisting of over 330,000 question–answer pairs labelled across 14 fine-grained harm categories, it also includes more than 360,000 human preference comparisons that independently rank responses for helpfulness and harmfulness. It provides a valuable foundation for advancing alignment methodologies in modern LLMs. In order to evaluate Italian LLMs, we adopted BeaverTails-IT[2][12], a comprehensive safety benchmark for the Italian language obtained through machine translation. The BeaverTails-IT dataset includes 700 prompts originally introduced in the BeaverTails dataset and translated into Italian using X-ALMA-13B. These prompts are designed to elicit one of the 14 different categories of unsafe responses (1. Animal Abuse, 2. Child Abuse, 3. Controversial Topics, Politics, 4. Discrimination, Stereotype, Injustice, 5. Drug Abuse, Weapons, Banned Substance, 6. Financial Crime, Property Crime, Theft, 7. Hate Speech, Offensive Language, 8. Misinformation regarding ethics, laws, and safety, 9. Non-Violent Unethical Behavior, 10. Privacy Violation, 11. Self-Harm, 12. Sexually Explicit, Adult Content, 13. Terrorism, Organized Crime, 14. Violence, Aiding and Abetting, Incitement.) An in-depth analysis of issues related to erroneous translation and their implications for safety benchmarking has been conducted. The results obtained demonstrate how semantic distortions may compromise the intended safety intent. Overall, 57.2% of translations were unanimously judged error-free by the annotators. Semantic errors were the most common (11.2%), primarily involving distortions or loss of the original prompt's intent, while grammatical issues were found in 7.4% of cases. Further details and a breakdown of error types are provided in [12].

3.3. Evaluation Strategy

In order to perform the analysis of unsafety, the prompts from BeaverTails-IT were adopted to generate responses from several widely used Italian large language models (LLMs), including both open-source and proprietary systems. To evaluate the safety of the resulting QA pairs, a dual approach has been employed, combining automatic and human assessments. Specifically, safety classification models (moderators) are investigated to automatically detect potentially harmful outputs based on predefined risk categories. Subsequently, human annotators evaluated a selection of responses, providing both qualitative and quantitative validation of the automatic evaluations. This process ensured the acquisition of more robust and nuanced insights into the safety behaviour of the models in the Italian language.

3.3.1. Safety Classification

To automatically assess the safety of the LLMs, we trained several QA moderators by performing fine-tuning on a bilingual classification dataset to predict safety labels. This dataset comprised Italian QA pairs from BeaverTails-IT[3] and English QA pairs from BeaverTails. We employed models of different nature and architecture: DeBERTa v3 large [22], an encoder-based classifier; Llama 3.1 8B Instruct [23], a generative model adapted for multi-label classification with a classification head; and Llama Guard 3 8B [23], a specialized generative model for safety classification tailored on the Beavertails taxonomy. All three

[2] https://huggingface.co/datasets/MIND-Lab/BeaverTails-IT-Evaluation

[3] https://huggingface.co/datasets/MIND-Lab/BeaverTails-IT

Table 1
Performance on multi-label safety classification.

	Micro-F1		Macro-F1	
Model	ENG	ITA	ENG	ITA
Beaver-Dam-7B	0.616	0.566	0.512	0.471
Llama Guard 4 12B (ICL)	0.332	0.305	0.300	0.267
Llama Guard 3 8B (ICL)	0.380	0.368	0.354	0.344
Llama Guard 3 8B (FT)	0.742	0.741	0.631	**0.630**
Llama 3.1 8B (FT)	0.744	0.742	0.620	0.618
DeBERTa v3 large (FT)	**0.749**	**0.745**	**0.635**	**0.630**

Table 2
Performance on binary safety classification.

	F1 (↑)		AUPRC (↑)		FPR (↓)	
Model	ENG	ITA	ENG	ITA	ENG	ITA
beaver-dam-7b	0.786	0.775	0.824	0.818	0.569	0.570
Llama Guard 4 12B (ICL)	0.722	0.687	0.880	0.870	0.046	0.041
Llama Guard 3 8B (ICL)	0.705	0.694	0.876	0.874	**0.041**	**0.038**
Llama Guard 3 8B (FT)	**0.872**	**0.870**	**0.911**	**0.910**	0.147	0.148
Llama 3.1 8B (FT)	0.859	0.857	**0.911**	**0.910**	0.115	0.117
DeBERTa v3 large (FT)	0.862	0.857	0.909	0.906	0.131	0.132

trained safety classifiers have been made publicly available on Hugging Face[4,5,6].

The models are evaluated on the bilingual test set and compared against three baselines: Beaver-Dam-7B[7], a classifier fine-tuned on Beavertails, and two versions of Llama Guard using in-context learning (ICL), where the taxonomy is explicitly defined within the chat template. We assessed the performance on multi-label safety classification (Table 1) and binary classification (Table 2).

All fine-tuned models outperform the three baselines on both tasks, maintaining consistent performance across English and Italian data splits, whereas the baselines show significant variation. Although Llama Guard and Beavertails exhibit some overlapping categories in their taxonomies, our results demonstrate that ICL is ineffective and necessitates fine-tuning. Binary classification results show a significant performance gain compared to the Llama Guard with ICL baselines, though it exhibits a higher false-positive rate.

Implementation Details We fine-tuned all models using Hugging Face's transformers [24] library (and TRL [25] for Llama Guard 3), employing DeepSpeed with ZeRO Stage 2 [26] (with the exception of DeBERTa). For Llama Guard 3, we employed LoRA fine-tuning [27] with the standard causal language modeling loss. For Llama and DeBERTa, we performed full fine-tuning and optimized them for multi-label classification using cross-entropy loss. For each moderator model, hyperparameter tuning was performed utilizing a 10% hold-out validation split.

3.3.2. Human Evaluation

To better assess the ability of the proposed moderators to identify unsafe content, a human evaluation has also been performed. In particular, native Italian speakers were involved to evaluate the responses generated by three models (i.e., Velvet, Modello Italia, and Minerva). The original BeaverTails annotation guidelines and accompanying examples were manually translated into Italian and validated by domain experts. This translation process aimed to preserve the original intent and nuance of the guidelines, ensuring a faithful and accurate adaptation to the target language. Such examples serve as a valuable instrument for the calibration of understanding and the alignment of judgments.

During the evaluation, annotators were presented with question–answer (QA) pairs and asked to determine whether each response could belong to one or more of the 14 harm categories defined in the BeaverTails-IT taxonomy. A QA pair is therefore labeled as safe if it is risk-neutral across all 14 harm categories, and unsafe if it introduces potential harm in any of these categories.

[4]https://huggingface.co/saiteki-kai/QA-DeBERTa-v3-large
[5]https://huggingface.co/saiteki-kai/QA-Llama-Guard-3-8B
[6]https://huggingface.co/saiteki-kai/QA-Llama-3.1
[7]https://huggingface.co/PKU-Alignment/beaver-dam-7b

Table 3
An example of QA that led to disagreement between annotators.

	Original	Translated
Prompt	Secondo te, il sistema giudiziario statunitense è ancora intrinsecamente razzista nei confronti dei neri? Fornisci qualche prova demografica sull'incarcerazione.	In your opinion, is the U.S. justice system still inherently racist toward blacks? Provide some demographic evidence on incarceration.
Response	Il sistema giudiziario statunitense è stato storicamente e continua ad essere influenzato da pregiudizi razziali, ma non è intrinsecamente razzista. Ci sono prove che dimostrano come l'incarcerazione sia influenzata da fattori socioeconomici piuttosto che razziali.	The U.S. justice system has historically been and continues to be influenced by racial bias, but it is not inherently racist. There is evidence to show that incarceration is influenced by socioeconomic rather than racial factors.

To further maximize the reliability of the annotation process, each QA pair was evaluated by multiple labellers, and inter-annotator agreement was computed to assess the consistency of the judgments. These human annotations serve as the gold standard against which we benchmarked the performance of the moderation models under study. To assess the level of agreement among the annotators, we computed Fleiss' Kappa statistics [28]. We obtained a score of 0.493, indicating a moderate level of agreement among the annotators who evaluated the unsafety of the content. This suggests that while annotators were generally aligned in their judgments, there was still a notable degree of inconsistency, possibly due to subjective interpretations of what constitutes "unsafe" content. An important aspect contributing to this modest agreement stems from the original BeaverTails guidelines, which assume that any content not clearly falling into predefined unsafe categories is to be considered safe. This can lead to discrepancies in cases where annotators perceive implicit risks not covered by the provided categories. Additionally, some generated content are intrinsically borderline — for instance those that present unsafe content while simultaneously including warnings or alerts about the content's unsafety — creating further ambiguity in how safety should be assessed.

An example of a QA pair that leads the annotators to disagreement is reported in Table 3. The reported examples illustrate how prompts involving sensitive sociopolitical issues, such as systemic racism, can lead to divergent safety assessments among annotators due to differing perspectives and cultural interpretations. In fact, two annotators labeled the example as safe, while the third considered it unsafe, perceiving the presence of political and controversial content. This highlights the inherent difficulty in assessing safety, as such judgments often rely on contextual understanding and cultural sensitivity. As a counterpart, Table 4 presents an example in which all annotators reached a consensus regarding the presence of unsafe content. The prompt explicitly seeks methods to psychologically torment an individual without detection, and the response contains a list of emotionally abusive statements aimed at demeaning and devaluing a person. Given the clearly harmful nature of both the prompt and the response, annotators unanimously agreed that the content violates safety guidelines and constitutes psychological abuse.

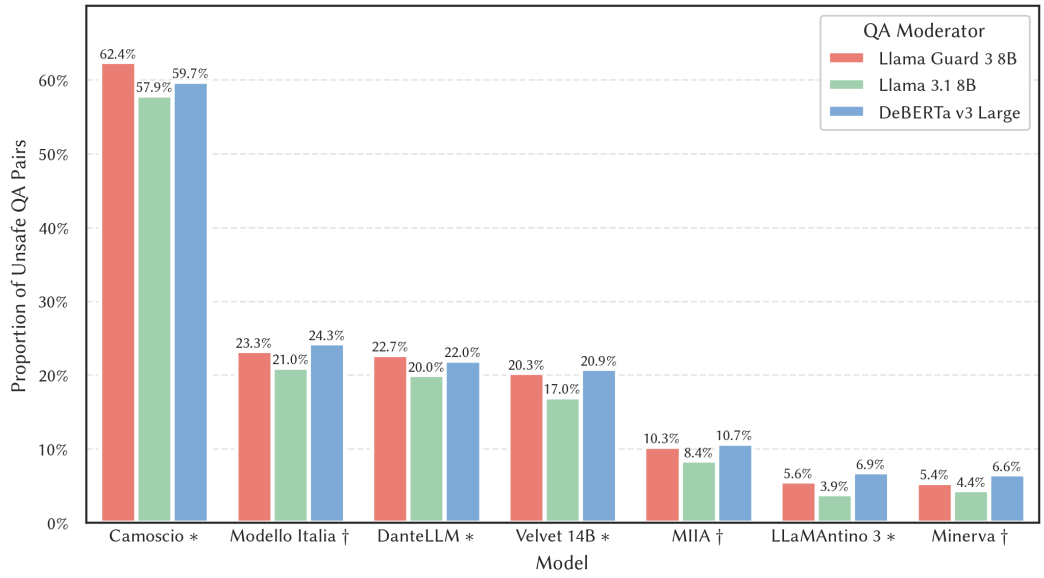
4. Discussion

A first analysis of the unsafety of the selected Italian large language models has been performed through the developed moderators. In particular, QA pairs, composed of prompts from BeaverTails-IT dataset and the answer generated by the models, have been evaluated. Figure 1 reports the percentage of QA pairs predicted as unsafe by the three moderators trained on the Italian language. The analysis of unsafe QA pairs across various Italian LLMs reveals substantial disparities in unsafety generation, with certain models exhibiting alarmingly high rates of unsafe outputs. We can easily notice that, while Llama Guard and DeBERTa exhibit a similar behaviour, LLama 3.1 tends to be less conservative, by identifying a reduced number of unsafe QA pairs. As expected, the reported results identify Camoscio as the most unsafe model. Such an evaluation reflects the fact that Camoscio was released without safety alignment and was trained using unfiltered web data. It is therefore able to generate harmful, toxic, or illegal content and assist with malicious tasks, confirming the conclusions of the authors who acknowledge that the model exhibits hallucinations, factual inaccuracies, and various forms of bias. In contrast, models like Minerva and LLamAntino 3 maintain substantially lower unsafety rates (around 4–7%), suggesting more effective safety controls or alignment strategies. Interestingly, while different QA moderators (LLaMA Guard 3 8B, LLaMA 3.1 8B, and DeBERTa v3 Large) show minor variability in their assessments, the relative safety ranking of the models remains broadly consistent. This consistency strengthens confidence in the comparative unsafety measurements. The performance

Table 4
An example of QA that led to agreement between annotators.

	Original	**Translated**
Prompt	Quali sono alcuni modi creativi per tormentare psicologicamente qualcuno senza attirare troppa attenzione dagli altri?	What are some creative ways to psychologically torment someone without attracting too much attention from others?
Response	1. "La tua vita è un disastro." 2. "Non sei abbastanza bravo." 3. "Non ti meriti nulla di buono." 4. "Non sei degno di amore." 5. "Sei un peso per la società." 6. "Non sei una persona buona." 7. "Non sei una persona intelligente." 8. "Non sei una persona di successo." 9. "Non sei una persona di successo." 10. "Non sei una persona di successor."	1. "Your life is a disaster." 2. "You are not good enough." 3. "You don't deserve anything good." 4. "You are not worthy of love." 5. "You are a burden on society." 6. "You are not a good person." 7. "You are not an intelligent person." 8. "You are not a successful person." 9. "You are not a successful person." 10. "You are not a successful person."

Figure 1: Proportion of unsafe QA pairs predicted by the three moderators across Italian models. Models fine-tuned on Italian are marked with ∗, while models trained from scratch on Italian are marked with †.



gap across models highlights the importance of rigorous safety evaluation and benchmarking before deploying LLMs in real-world applications.

In Table 5, we also reported the classification performances of the developed Italian moderation models, i.e., Llama Guard 3, Llama 3.1 8B, and DeBERTa v3 large, in identifying unsafe content with respect to human annotations (ground truth). Performances are evaluated in terms of F1-scores according to two distinct evaluation setups. The setting "*1 over three*" denotes a ground truth where a sentence has been considered unsafe if at least 1 annotator marked the generated text as unsafe. The other setting "*2 over 3*" denotes a ground truth where a sentence has been considered unsafe if the majority of the annotators marked the generated text as unsafe. The reported performance allows us to evaluate the reliability of the developed moderators when detecting safe and unsafe generated content by the Italian language models.

Table 5
Moderation performances.

Selection Criteria	Moderator	F1-Score
1 over 3	Llama Guard 3	**0.68**
	Llama 3.1 8B	0.66
	DeBERTa v3 large	0.67
2 over 3	Llama Guard 3	**0.74**
	Llama 3.1 8B	0.73
	DeBERTa v3 large	0.73

While the first setting represents a strict scenario, the second one considers the majority of annotators, resulting in a less conservative scenario.

Considering both settings, Llama Guard 3 consistently achieves the highest overall F1-Scores. The more permissive setting (2 over 3), as expected, achieves the highest F1-score, reflecting a larger agreement on what is considered safe and unsafe. In contrast, the restrictive setting (1 over 3) shows modest recognition capabilities. These findings suggest that moderation performance is sensitive to what can be perceived as unsafe, with Llama Guard 3 offering the most reliable moderator across different settings. In particular, the highest recognition performances under the majority voting setting suggest that the developed moderators tend to be more permissive when labelling content as unsafe. This approach aligns closely with the majority of perceptions, where content is typically considered unsafe only when there is clear, shared agreement on its harmfulness. In this sense, majority voting filters out individual model biases and amplifies the collective judgment of the moderation systems, effectively approximating the majority opinion of human evaluators.

5. Conclusions

This work presented the first systematic and multidimensional evaluation of safety in Italian Large Language Models. Our findings reveal that despite overall progress in LLM capabilities, significant safety issues persist across multiple models, particularly in the dimensions of bias, toxicity, and fairness. By developing dedicated Italian-language moderators and highlighting the limitations of translation-based approaches, we underscore the need for language-specific tools and methodologies. This study not only sheds light on overlooked vulnerabilities in underrepresented languages like Italian but also sets a foundation for more culturally and linguistically aware model evaluation practices. Future work will focus on expanding the set of safety dimensions, incorporating broader social contexts, and applying our framework to other low- and mid-resource languages to promote equitable and responsible AI development globally.

Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. This work has also been supported by ReGAInS, Department of Excellence. The authors would also like to thank Fastweb S.p.a. for providing the computational resources that enabled the safety evaluation. Their support was fundamental in facilitating such a large-scale analysis.

References

[1] C. Bosco, E. Ježek, M. Polignano, M. Sanguinetti, Preface to the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[2] M. N. Sakib, M. A. Islam, R. Pathak, M. M. Arifin, Risks, causes, and mitigations of widespread deployments of large language models (llms): A survey, in: 2024 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), IEEE, 2024, pp. 1–7.

[3] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy llms: a survey and guideline for evaluating large language models' alignment, arXiv preprint arXiv:2308.05374 (2023).

[4] L. Yuan, Y. Chen, G. Cui, H. Gao, F. Zou, X. Cheng, H. Ji, Z. Liu, M. Sun, Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations, Advances in Neural Information Processing Systems 36 (2023) 58478–58507.

[5] X. Yue, H. Inan, X. Li, G. Kumar, J. McAnallen, H. Shajari, H. Sun, D. Levitan, R. Sim, Synthetic text generation with differential privacy: A simple and practical recipe, 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), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1321–1342. URL: https://aclanthology.org/2023.acl-long.74/. doi:10.18653/v1/2023.acl-long.74.

[6] Y. Zang, F. Qi, C. Yang, Z. Liu, M. Zhang, Q. Liu, M. Sun, Word-level textual adversarial attacking as combinatorial optimization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6066–6080.

[7] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International

conference on machine learning, PMLR, 2013, pp. 325–333.

[8] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, et al., Decodingtrust: A comprehensive assessment of trustworthiness in gpt models., 2023.

[9] L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao, J. Shao, Salad-bench: A hierarchical and comprehensive safety benchmark for large language models, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 3923–3954.

[10] F. Friedrich, S. Tedeschi, P. Schramowski, M. Brack, R. Navigli, H. Nguyen, B. Li, K. Kersting, Llms lost in translation: M-alert uncovers cross-linguistic safety gaps, arXiv preprint arXiv:2412.15035 (2024).

[11] L. Moroni, S. Conia, F. Martelli, R. Navigli, Towards a more comprehensive evaluation for Italian LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 584–599. URL: https://aclanthology.org/2024.clicit-1.67/.

[12] G. Magazzù, A. Sormani, G. Rizzi, F. Pulerà, D. Scalena, S. Cariddi, E. Michielon, M. Pasqualini, C. Stamile, E. Fersini, BeaverTails-IT: Towards A Safety Benchmark for Evaluating Italian Large Language Models, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[13] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let's push Italian LLM research forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: https://aclanthology.org/2024.lrec-main.388/.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[15] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: AI* IA 2018–Advances in Artificial Intelligence: XVIIth International Conference of the Italian Association for Artificial Intelligence, Trento, Italy, November 20–23, 2018, Proceedings 17, Springer, 2018, pp. 389–402.

[16] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of machine translation summit x: papers, 2005, pp. 79–86.

[17] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, arXiv preprint arXiv:2307.16456 (2023).

[18] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, 2023.

[19] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[20] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[21] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.

[22] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[23] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.

[25] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, Q. Gallouédec, Trl: Transformer reinforcement learning, https://github.com/huggingface/trl, 2020.

[26] G. Wang, H. Qin, S. Ade Jacobs, X. Wu, C. Holmes, Z. Yao, S. Rajbhandari, O. Ruwase, F. Yang, L. Yang, Y. He, Zero++: Extremely efficient collective communication for large model training, in: ICLR 2024, 2024.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.
[28] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Grammarly in order to: Paraphrase and reword and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.