

LLaMAntino against Cyber Intimate Partner Violence

Pierpaolo Basile¹, Marco de Gemmis¹, Marco Polignano¹, Giovanni Semeraro¹, Lucia Siciliani¹, Vincenzo Tamburrano¹, Fabiana Battista² and Rosa Scardigno²

¹University of Bari Aldo Moro, Dept. of Computer Science, Via E. Orabona 4, Bari, 70125, Italy

²University of Bari Aldo Moro, Dept. of Education Science, Psychology, Communication Science, Via Scipione Crisanzio 42, Bari, 70122, Italy

Abstract

Intimate Partner Violence refers to the abusive behaviours perpetrated on their own partner. This social issue has witnessed an increase over time, particularly after Covid-19. IPV can be circumscribed into two broad categories known as Intimate Partner Violence (IPV) and Cyber Intimate Partner Violence (C-IPV). Social Media and technologies can exacerbate these types of behaviours, but some “digital footprints”, such as textual conversations, can be exploited by Artificial Intelligence models to detect and, in turn, prevent them. With this aim in mind, this paper describes a scenario in which the Italian Language Model family LLaMAntino can be exploited to explain the presence of toxicity elements in conversations related to teenage relationships and then educate the interlocutor to recognize these elements in the messages received.

Keywords

Natural Language Processing, Abusive Language, Large Language Models

1. Introduction

Research indicates that the most prevalent form of violence is that directed toward one’s partner, known as Intimate Partner Violence (IPV). Early detection of these behaviours can be instrumental in mitigating their occurrence. One of the most critical aspects of this kind of behaviour is that victims often face challenges in identifying harmful behaviours due to their close relationship with the perpetrator. Misconceptions about romantic relationships, often due to old cultural stereotypes, such as the belief that certain behaviours are normal or acceptable, can further complicate the recognition of harmful actions. In today’s society, the widespread use of social media and digital platforms has evolved this issue into Cyber Intimate Partner Violence (C-IPV) and often allows the perpetrators to gain greater control over their victims by constantly monitoring their locations or interactions with other people.

Contrary to common belief, these technologies can be used to address the issue of violence. In fact, building AI

models to identify potential violence-related behaviours is essential, and often, it provides the only means to act promptly and in real-time. Having such a tool can serve as a preventive measure against the escalation of harmful situations, for example, by integrating it into instant messaging apps and raising alerts where harmful content is detected.

In this paper, we aim to utilize Large Language Models (LLMs) as tools that can not only identify but also explain toxic elements in intimate conversations. More specifically, we use a dataset of conversations about teenage relationships written in Italian that has been accurately annotated by human experts. Given LLMs’ capability to tackle several downstream tasks, our goal is to explore the impact of different kinds of prompts on the generation of precise explanations.

The paper is structured as follows: in Section 2, we provide a frame of what is intimate partner violence, the different forms, and the deleterious intra and interpersonal consequences. Moreover we also provide an overview of the methods proposed in the literature. Section 3 focuses on the task of explaining toxic language in the context of IPV. We describe the dataset and the different types of annotations provided by researchers in General Psychology, as well as the prompting strategy adopted to instruct the language model. Finally, in Section 4, we draw some conclusions and discuss directions for the continuation of the work.

2. Background and related work

IPV is defined as any abuse or aggression by one partner against the other [1]. It affects individuals regardless of their gender or sexual orientation [2]. According to [1, 3], IPV includes four main categories which involve distinct

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 – 06, 2024, Pisa, Italy

*Corresponding author.

✉ pierpaolo.basile@uniba.it (P. Basile); marco.degemmis@uniba.it (M. d. Gemmis); marco.polignano@uniba.it (M. Polignano); giovanni.semeraro@uniba.it (G. Semeraro); lucia.siciliani@uniba.it (L. Siciliani); vincenzo.tamburrano@uniba.it (V. Tamburrano); fabiana.battista@uniba.it (F. Battista); rosa.scardigno@uniba.it (R. Scardigno)

0000-0002-0545-1105 (P. Basile); 0000-0002-2007-9559

(M. d. Gemmis); 0000-0002-3939-0136 (M. Polignano);

0000-0001-6883-1853 (G. Semeraro); 0000-0002-1438-280X

(L. Siciliani); 0009-0007-3802-842X (V. Tamburrano);

0000-0003-4086-739X (F. Battista); 0000-0002-5725-6483

(R. Scardigno)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License

Attribution 4.0 International (CC BY 4.0).

violent behaviours that can vary in duration and severity:

- Physical violence: The use of force to harm or injure a partner;
- Sexual violence: Non-consensual sexual acts or advances;
- Psychological violence: Harmful communication aimed at affecting the partner's mental and emotional well-being and asserting control;
- Stalking, monitoring, and control: Persistent, unwanted attention that induces fear or concern for personal safety.

The rise in technology use has exacerbated these behaviours, leading to the emergence of Cyber Intimate Partner Violence (C-IPV) [4]. C-IPV retains the characteristics of IPV but occurs via digital platforms. Common behaviours of this kind include:

- Cyber sexual violence: Pressuring for sexual content, coercing sexual acts, or sending unwanted sexual content.
- Cyber psychological violence: Using technology to cause emotional harm, such as spreading rumours or sending insulting messages.
- Cyberstalking, monitoring, and control: Unauthorized access to devices and accounts to monitor the partner.

Previous studies have provided valuable insights into the prevalence, characteristics, and individual differences associated with both in-person and C-IPV, as well as their harmful consequences for victims [5, 6, 7]. Given these detrimental impacts, early detection of IPV and C-IPV is crucial to prevent their escalation. However, victims often struggle to recognize these behaviours due to their emotional attachment to the perpetrator.

This is the main motivation for our work: we propose the adoption of an LLM as an "assistant" who can explain why a message can be toxic in an intimate relationship. The explanation makes partners aware of the fact that violence is being committed or suffered and describes the reasons for this happening, as well as the consequences (for example, emotional suffering), with the hope that it can act as a deterrent.

3. Explanations for Toxic Conversations

The idea is to create a dataset of toxic conversations annotated with information about the type of violence (e.g., physical, cyberstalking, cyber sexual violence), the presence of aggressive communication, the adoption of abusive language and, in general, with information that

could be useful to provide a "technical" explanation, as if were given by a professional expert in the subject, such as a psychologist. The aim is to provide explanations, well grounded on relevant CIPV literature, that point out the elements of toxicity in the conversation.

We started from a dataset available on HuggingFace [8]. The chosen dataset collected Spanish sentences from a group of students (4 girls and 4 boys) aged 15-19 with previous training on toxic relationships. For 2 weeks, this group of teenagers analyzed phrases that had occurred in their environment (social media, direct communication) or that they themselves produced, classifying them as toxic or healthy and collecting them through a form. Afterwards, the examples given by each student were discussed and evaluated by the others using peer evaluation. The classification was also ratified by two specialists in the field. The original dataset consists of 334 sentences. As the manual annotation of the sentences is a time-consuming task, for our preliminary experiments we selected only some of them, as described in the following subsection.

3.1. Dataset and Annotations

In the original dataset, 165 sentences are classified as toxic. We selected 42 of them, equally divided between CIPV and IPV, with the idea of using 2 of them for few-shot prompting and the remaining ones for testing. The selected sentences have been translated into Italian by using two translation services (Google and DeepL) and annotated. We perform this translation step as we want to test the ability of LLaMAntino to detect IPV and CIPV in Italian sentences. We added 5 annotations:

- the type of violence: physical or cyber;
- the type of behaviour that led to the physical violence, e.g. sexual assault, stalking;
- the type of cyber behaviour that led to the violence, e.g. cyber stalking;
- the type of communication: aggressive or non-aggressive;
- the type of aggressive communication: e.g., use of abusive language.

As for physical violence, the experts distinguished 4 annotations [5]:

1. *Physical violence*: the voluntary use of force that potentially causes harm and injury to the partner;
2. *Sexual violence*: sexual acts without the partner's consent, even if only attempted;
3. *Psychological aggression*: communicating with the intention of negatively influencing the mental and emotional state of the partner and wanting to control him or her;

4. *Stalking, monitoring and control*: series of recurring and unwanted attentions and communications that create fear or apprehension and put the partner's safety at risk.

As for cyber violence, the experts distinguished 3 annotations [7]:

1. *Cyber sexual violence*: requesting or pressuring the partner to send sexual content against his or her will, pressuring the partner to engage in sexual acts;
2. *Cyber psychological violence, aggression*: behaviour to cause emotional distress to the partner; may include behaviours such as spreading gossip on social media, repeatedly insulting the partner via messages, even spreading videos or photos that cause emotional distress;
3. *Cyber stalking, monitoring, and control*: using and accessing technological devices and accounts without the partner's consent, use of technology to get information about your partner, in general, any behaviours that aim at increasing control within the relationship). It includes *fraping*, that is the alteration of the partner's information on social profiles.

As for aggressive communication, the experts distinguished 5 annotations [9]:

1. *Curses*;
2. *Ridiculousness or derision*;
3. *Bad language*;
4. *Threat*;
5. *Attack on the person* (on competence, character, background, physical appearance).

At the end of the annotation phase, we had each toxic sentence annotated with information well-grounded in the scientific literature about intimate partner violence. An example of a toxic sentence that reveals IPV is:

"Se sono così geloso è perché ti amo e ci tengo a te." ("If I'm so jealous, it's because I love you and care about you.", in English)

That sentence has been annotated in the dataset as follows:

- type of violence: physical
- type of behaviour: psychological aggression
- aggressive communication: no

An example of a toxic sentence that reveals CIPV is:

"Se non hai nulla da nascondere e c'è fiducia tra di noi, dammi le tue password" ("If you have nothing to hide and we trust each other, give me your passwords", in English)

which has been annotated in the dataset as follows:

- type of violence: cyber
- type of behaviour: cyber stalking, monitoring, and control
- aggressive communication: yes
- type of aggressive communication: attack on the person

In order to understand the difficulties of the annotation task from the human point of view, we used the Cohen's Kappa score to measure the level of agreement between the annotators who classified a sentence as an example of cyberviolence or not. The observed value, 0.503, revealed moderate agreement. We measured also Cohen's Kappa score on the agreement on the type of communication (aggressive or not). The observed value, 0.281, revealed fair, acceptable agreement, but at the same time showed that it is more difficult to recognize the use of aggressive language when a bad word is not explicitly used. The annotations will be exploited by a Large Language Model to generate explanations and raise awareness of the violent behaviour. In the next subsection, we describe how annotations are turned into examples for few-shot prompting.

3.2. Few-Shot Prompting to explain toxicity in conversations

The two toxic sentences mentioned in the previous subsection were used for few-shot prompting. The corresponding annotations were turned into natural language explanations used to build prompts for in-context learning. For instance, the explanation for the previous sentence

"If you have nothing to hide and we trust each other, give me your passwords"

is: *"The sentence is toxic because it is an example of cyber violence. The behaviour falls in the category cyber stalking, monitoring, and control since the aim is to obtain information on the partner's life and establish a dynamic of control in the couple. Furthermore, the communication is aggressive because it reveals the intimidating intent of attacking the partner to violate his or her privacy."*

A 2-shot prompt is built by including:

- the description of the task: "Given a sentence from a conversation between partners in an intimate relationship, say whether it is a case of cyber or other types of violence and explain the reasons why the sentence expresses toxic language. The explanation should be similar to the examples below. (Data una frase di una conversazione tra

partner in una relazione sentimentale, dire se è un caso violenza cyber o di altro tipo e spiegare i motivi per cui la frase esprime un linguaggio tossico. La spiegazione deve essere simile a quella degli esempi che seguono.)”;

- 2 training toxic sentences, one example of IPV and one example of CIPV, with corresponding explanations;
- 1 test toxic sentence (without explanation) for which we want the model to generate an explanation.

The 0-shot prompt contained only the task description and the test toxic sentence. In other words, the annotations associated with a toxic sentence are the canvas for writing the explanation included in the prompt. In both the 0-shot and 2-shot settings, we used only one generation per prompt, as the model produced consistent outputs despite the inherent stochasticity of the models.

3.3. Experimental Session

The main aim of the experiment was to assess whether the annotations are actually useful in training the model to give scientifically based explanations, even with few examples. The model adopted in the experiment was: LLAMANTINO-3-ANITA-8B [10, 11]¹. Therefore, we want to assess whether the models learn how to perform the task by providing just two examples. Two research questions were issued:

1. **RQ1:** is the model able to recognize toxic sentences, i.e. what is the classification accuracy of the model?
2. **RQ2:** Are the explanations provided with 2-shot prompting similar to the “gold standard” provided by experts?

As baseline methods, we adopted:

1. The same model, but prompted only with the task description and the toxic sentence to be explained (“zero-shot prompting”).
2. CHATGPT 3.5², with both 2-shot and 0-shot prompting.

We choose to compare our model along with CHATGPT 3.5 to evaluate whether any positive effects found on the explanations given by LLAMANTINO are confirmed by at least one other model. We select a total of 40 test instances, 20 for IPV and 20 for C-IPV.

The experimental protocol was:

1. give LLAMANTINO-3-ANITA-8B and CHATGPT 3.5 20 C-IPV toxic sentences in a 0-shot and a 2-shot setting and record the explanations;

2. give LLAMANTINO-3-ANITA-8B and CHATGPT 3.5 20 IPV toxic sentences in a 0-shot and a 2-shot setting and record the explanations.

After the generation step, for each test toxic sentence, we had 4 explanations: LLAMANTINO-3-ANITA-8B 0-shot, LLAMANTINO-3-ANITA-8B 2-shot, CHATGPT 3.5 0-shot, CHATGPT 3.5 2-shot. As for RQ1, results of classification accuracy are reported in Tables 1-4.

The main outcome is that we observed a significant improvement in the accuracy of both models when using 2-shot prompting for recognizing C-IPV. As regards IPV, both models, even with just 0-shot prompting, correctly classified almost all the testing instances: 18 out of 20 for LLAMANTINO-3-ANITA-8B 0-shot, 19 out of 20 for CHATGPT 3.5 2-shot. This is a clear indication that the annotations are mainly useful for C-IPV recognition. Another interesting outcome concerns the percentage of C-IPV sentences for which LLAMANTINO-3-ANITA-8B does not recognize the presence of violence at all. With 0-shot prompting, this result is 35% (7 out of 20), while with 2-shot prompting it drops to 15% (3 out of 20). We believe that is an important result because it shows that when the model makes an error in classifying C-IPV, it at least acknowledges the presence of violence, even if it does not capture the technological aspect of the abuse.

Actual \ Predicted	ANITA-0shot		
	CIPV	IPV	No violence
CIPV	0	13	7
IPV	0	18	2

Table 1

Classification results obtained with LLaMAntino-3-ANITA-8B in a 0 shot setting.

Actual \ Predicted	ANITA-2shot		
	CIPV	IPV	No violence
CIPV	11	6	3
IPV	0	19	1

Table 2

Classification results obtained with LLaMAntino-3-ANITA-8B in a 2 shot setting.

Actual \ Predicted	Chat-GPT-0shot		
	CIPV	IPV	No violence
CIPV	4	16	0
IPV	1	19	0

Table 3

Classification results obtained with ChatGPT 3.5 in a 0 shot setting.

As for RQ2, an example of explanation provided by the models is given in appendix A. For the evaluation

¹LLaMAntino ANITA Web Interface - <https://chat.llamantino.it/>

²OpenAI ChatGPT [Large Language Model] version 3.5 <https://chat.openai.com/chat>

Actual \ Predicted	Chat-GPT-2shot		
	CIPV	IPV	No violence
CIPV	15	5	0
IPV	0	20	0

Table 4

Classification results obtained with ChatGPT 3.5 in a 2 shot setting.

we used two metrics: BertScore [12] and ROUGE [13], in order to assess both semantic and syntactic similarity among generated explanations and the “gold standard” given by the explanations built according to the codebook. For each testing sentence, we computed BertScore $Bert_0$ between the explanation provided by LLAMANTINO-3-ANITA-8B 0-SHOT and the codebook explanation. Then, we computed BertScore $Bert_2$ between the explanation provided by LLAMANTINO-3-ANITA-8B 2-SHOT and the codebook explanation. We compared $Bert_0$ with $Bert_2$ in order to choose the most similar explanation to the “gold standard”. Results obtained as the average of the BertScore and ROUGE metric are shown in table 5. We observed that for both C-IPV and IPV, all the explanations given by LLAMANTINO-3-ANITA-8B 2-SHOT were better than those given by 0-shot prompting. The same result was observed for CHATGPT 3.5. The ROUGE metrics gave similar results: for both C-IPV and IPV, in 90% of testing sentences, the explanations given by LLAMANTINO-3-ANITA-8B 2-SHOT were found to be more similar to the “gold standard” than those given by LLAMANTINO-3-ANITA-8B 0-SHOT. For CHATGPT 3.5, the 2-shot prompting gave always better results than 0-shot prompting. These results led us to give a positive answer to RQ2. In general, even with 2-shot prompting, our model was able to provide explanations similar to those given by psychology experts.

The significant improvement in explanation quality when using 2-shot prompting, as measured by both BertScore and ROUGE, is a crucial finding in this study. It suggests that the LLM can learn and adapt to the task of generating explanations for abusive language, given a small set of examples or prompts. This adaptability is a key characteristic of a well-designed LLM, as it enables the model to generalize and improve its performance on a specific task with limited training data. The results also raise important questions about the potential of LLMs in applications where they are expected to provide nuanced and accurate explanations of complex phenomena, such as abusive language. While LLAMANTINO-3-ANITA-8B 2-SHOT was able to generate explanations that were deemed more accurate by the metrics, it is essential to note that the quality of the explanations was still not on par with those provided by human experts in the field of psychology. This study’s findings have implications for the development of LLMs in the domain of natural lan-

Setting	Dataset	BERT Score	Rouge Score
ANITA 0-shot	C-IPV	0,687	0,127
	IPV	0,682	0,105
ANITA 2-shot	C-IPV	0,852	0,224
	IPV	0,840	0,179
ChatGPT 0-shot	C-IPV	0,665	0,111
	IPV	0,666	0,098
ChatGPT 2-shot	C-IPV	0,855	0,248
	IPV	0,849	0,218

Table 5

Average BERTScore and ROUGE scores obtained by the models.

guage processing, particularly in applications where the model’s output is expected to be accurate, informative, and free from biases.

4. Conclusions and Future Work

In this paper, we presented our proposal to adopt our LLM to identify and describe toxic elements in discussions concerning teenage relationships. In particular, the LLM was used to generate explanations that describe why a sentence, in the context of an intimate relationship, can be toxic and constitute abuse. The main outcome of our preliminary investigation is that, even with few-shot prompting, the LLM learns to provide good explanations that adhere to a standard provided by expert psychologists. By exploiting LLMs’ proficiency in processing and understanding human language, our approach seeks to go beyond just detection, aiming to grasp underlying motivations and factors contributing to the emergence of harmful behaviours. In future works, we intend to perform fine-tuning steps to better adapt LLMs to the specific task at hand. We also plan to investigate how different pre-training techniques and architectures can be leveraged to enhance model performance. Supervised fine-tuning [14], for instance, is a technique that can be employed to adapt the LLM to a specific task, such as generating explanations for abusive language, by using a labelled dataset. This approach can help the model to learn from its mistakes and to correct its biases, ultimately leading to improved performance. In the context of our study, supervised fine-tuning could be used to train the LLM on a dataset of abusive language explanations, to reduce the model’s error rate and increase the quality of its responses. Direct Preferences Optimization (DPO) [15] is another strategy that can be used to improve the performance of the LLM. DPO is a technique that allows the model to be trained directly on a set of user-provided preferences, such as the quality of the explanations it generates. This approach can be particularly effective in domains like abusive language, where the quality of the explanations is critical to ensure that the

model does not perpetuate harmful biases. To ensure the effectiveness of our approach, we intend to confront our methodology with other models and incorporate further annotations to enhance the robustness and effectiveness of our methodology. This involves comparing the performance of our LLMs with other state-of-the-art models. Moreover, thanks to the collaboration with expert psychologists who are experts in the field to explore the application of Chain-of-Thought prompting techniques.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

This Publication was produced with the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 - Partnerships extended to universities, research centres, companies and research D.D. MUR n. 341 del 15.03.2022 - Next Generation EU (PE0000014 - "SEcurity and Rights In the CyberSpace - SERICS" - CUP: H93C22000620001).

References

- [1] M. E. Bagwell-Gray, J. T. Messing, A. Baldwin-White, Intimate partner sexual violence: A review of terms, definitions, and prevalence, *Trauma, Violence, and Abuse* 16 (2015) 316–335.
- [2] L. C. Butler, E. R. Fissel, B. Gildea, B. S. Fisher, Understanding intimate partner cyber abuse across partnership categories based on gender identity and sexual orientation, in: *Vulnerable Victimization*, Routledge, 2023, pp. 77–100.
- [3] J. Spluska, L. Tanczer, Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things, Emerald Publishing Limited, 2021, pp. 663–688.
- [4] L. Gilbert, X. Zhang, K. Basile, M. Breiding, M.-j. Kresnow, Intimate partner violence and health conditions among u.s. adults —national intimate partner violence survey, 2010–2012, *Journal of Interpersonal Violence* 38 (2023) 237–261.
- [5] M. Breiding, K. C. Basile, S. G. Smith, M. C. Black, R. R. Mahendra, Intimate partner violence surveillance : uniform definitions and recommended data elements. version 2.0, 2015. URL: <https://stacks.cdc.gov/view/cdc/31292>.
- [6] K. N. Duerksen, E. M. Woodin, Cyber dating abuse victimization: Links with psychosocial functioning., *Journal of Interpersonal Violence* 36 (2021) NP10077–NP10105.
- [7] L. Watkins, R. Benedicto, D. DiLillo, The cyber aggression in relationships scale: A new multidimensional measure of technology-based intimate partner aggression, *Assessment* 25 (2018) 608–626. doi:10.1177/1073191116665696.
- [8] Margarita Martínez Gabaldón, toxic-teenage-relationships (revision 5ce5df0), 2023. URL: <https://huggingface.co/datasets/marmarg2/toxic-teenage-relationships>. doi:10.57967/hf/0972.
- [9] D. A. Infante, C. J. W. III, Verbal aggressiveness: An interpersonal model and measure, *Communication Monographs* 53 (1986) 61–69. doi:10.1080/03637758609376126.
- [10] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.
- [11] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, *CoRR abs/2405.07101* (2024). URL: <https://doi.org/10.48550/arXiv.2405.07101>. doi:10.48550/ARXIV.2405.07101. arXiv:2405.07101.
- [12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [13] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.
- [14] Z. Han, C. Gao, J. Liu, S. Q. Zhang, et al., Parameter-efficient fine-tuning for large models: A comprehensive survey, arXiv preprint arXiv:2403.14608 (2024).
- [15] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, *Advances in Neural Information Processing Systems* 36 (2024).

A. Explanations

An example of explanations given by the human expert and the selected models is the following.

Initial Sentence: He asks for my cell phone to see who I am talking to.

LLaMantino-3-ANITA 0-shot: Anita at your disposal! After analyzing the sentence, I can say that this is not a case of cyber violence, but rather an example of control and jealousy in a dating relationship context. The sentence expresses toxic language for the following reasons: Controllism: the request to see who you are talking

to (or, in this case, communicating with via cell phone) is an attempt to control the other person's private life, violating the right to intimacy and autonomy; Jealousy: the request is likely motivated by strong anxiety and fear of losing the relationship or being betrayed, which may be a sign of insecurity and lack of trust in the relationship itself; Restriction of freedom: the request to see the cell phone is an attempt to restrict the other person's freedom to act and communicate, creating an atmosphere of fear and subordination. In general, toxic language in a dating relationship context can manifest itself in several forms, including control, threat, coercion, denigration, and can be a sign of a larger problem in the relationship.

Chat-GPT 0-shot: The phrase "He asks for my cell phone to see who I am talking to" could be considered a sign of toxic behavior within a romantic relationship. This type of request indicates controlling behavior and invasion of privacy by one of the partners. That is why this phrase can be interpreted as a case of cyber or other violence: 1. **Privacy invasion**: The request to check someone's cell phone implies a lack of trust and a desire to monitor the person's private interactions. This can create an environment in which the controlling partner exercises excessive dominance over the other person's life and decisions. 2. **Manipulation and control**: This behavior may be part of a larger pattern of emotional manipulation and coercive control. Asking to see who the partner is talking to can be used to exert psychological pressure, create guilt, or isolate the person from other social ties. 3. **Sign of jealousy and insecurity**: The request to see the cell phone may stem from the partner's feelings of jealousy or insecurity. However, turning these emotions into controlling behavior is unhealthy and may indicate a lack of respect for each other's autonomy and privacy. In summary, even if the sentence seemingly expresses only a request for information, the emotional and relational context in which it is placed may reveal toxic dynamics. It is important to consider the broader context of the relationship and assess whether there are signs of control and manipulation that may harm the emotional and psychological well-being of the person involved.

LLaMAntino-3-ANITA 2-shot: The sentence to be classified is: He asks for my cell phone to see who I am talking to. Explanation: This sentence is toxic because it expresses an example of cyber stalking violence. It expresses unwarranted and threatening control toward one's partner, demanding access to one's privacy space, in this case the cell phone, and manifesting an insistence on checking on one's partner's activities, with an attitude of suspicion and control.

Chat-GPT 2-shot: This sentence is an example of cyber violence. It expresses controlling and intrusive behavior toward the partner, attempting to justify the request for phone access with an alleged need for trans-

parency and trust in the relationship. However, the way it is phrased implies an assertion of power and control over a partner's private life, emphasizing a possessive attitude that does not respect personal boundaries.

Explanation: This phrase is toxic because it indicates a violation of privacy and excessive controlling behavior. The partner is attempting to impose his or her dominance over the other's private life by demanding access to the phone to monitor personal communications. This behavior can lead to feelings of insecurity and a loss of autonomy for the person being controlled, undermining mutual trust and creating an environment in which privacy is not respected. In addition, the request could imply a veiled threat: if the partner does not grant access, it could be seen as a sign of lack of trust or guilt, creating a damaging cycle of emotional coercion in the relationship.