

LLMs to Detect Cyber Child Abuse in the in Textual Conversations

Maria Teresa Baldassarre^{1,†}, Vita Santa Barletta^{1,†}, Vito Bavaro^{1,†}, Danilo Caivano^{1,†},
Alberto Pio De Matteis^{1,†}, Andrea Lippolis^{1,†}, Antonio Piccinno^{1,†}

¹ Department of Computer Science, University of Bari “Aldo Moro”, Via Orabona 4, 70121 Bari, Italy

Abstract

In contemporary online interactions, identifying inappropriate language and safeguarding minors from harmful communication is a critical challenge. This study explores the use of Large Language Models (LLMs) to analyze text, detecting patterns indicative of age-specific language and the presence of sexual or pornographic references. A fine-tuning of the LLaMAntino model was performed, using a dataset of synthetically generated sentences designed to replicate real-world scenarios. The fine-tuned model demonstrated enhanced performance compared to its baseline (given by LLaMAntino 3 ANITA 8B), providing detailed and context-sensitive explanations for its classifications.

The results highlight the potential of LLMs in addressing sensitive linguistic phenomena with precision, offering a foundation for detecting indirect combinations of sexual references in conversations involving minors. Future work can focus on incorporating real conversational data and involving subject matter experts to refine the model’s interpretability and reliability. Additionally, the exploration of advanced architectures and fine-tuning techniques will be considered to further balance model complexity and processing efficiency.

1. Introduction

In the modern digital age, the prevalence of harmful communication in online interactions has become an increasingly pressing issue, particularly when it comes to protecting minors. Inappropriate language, which may include age-improper expressions or sexually explicit content, poses significant risks to the emotional and psychological well-being of young individuals [1]. If left unaddressed, these dangerous interactions can contribute to unsafe digital environments, foster exploitative behaviours, and erode trust in online platforms. Identifying and mitigating such communication is crucial to ensuring a safer and more supportive space for minors in the digital landscape.

Modern Large Language Models (LLMs) offer a promising solution to this challenge due to their ability to analyze and interpret complex textual patterns and contexts. These advanced models,

Joint Proceedings of IS-EUD 2025: 10th International Symposium on End-User Development, 16-18 June 2025, Munich, Germany.

*Corresponding author.

†These authors contributed equally.

✉ mariateresa.baldassarre@uniba.it (M.T. Baldassarre); vita.barletta@uniba.it (V.S. Barletta);
v.bavaro4@studenti.uniba.it (V. Bavaro); danilo.caivano@uniba.it (D. Caivano); a.dematteis15@studenti.uniba.it
(A.P. De Matteis); a.lippolis34@studenti.uniba.it (Andrea Lippolis); antonio.piccinno@uniba.it (A. Piccinno)



Copyright © 2025 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

when fine-tuned with targeted datasets, can effectively identify sensitive linguistic phenomena and provide detailed, context-aware explanations. By leveraging the power of LLMs, researchers and developers can create tools to detect harmful communication patterns, safeguard vulnerable populations, and promote healthier digital interactions.

2. Related Work

Recent research on large language models (LLMs) in the field of harmful language detection and safeguarding minors has seen notable advancements. Studies have explored techniques for identifying harmful communication patterns, including predatory behavior, sexually explicit content, and age-inappropriate language. This involves fine-tuning models with domain-specific datasets and leveraging transfer learning to improve their ability to detect subtle linguistic cues indicative of harmful interactions, as demonstrated in [2,3].

Recent work has also focused on improving LLMs' capabilities to process multimodal inputs and apply advanced reasoning techniques for better contextual understanding. For instance, studies such as [4,5] have explored the integration of visual and textual data to enhance the detection of harmful content, including age-inappropriate material, in digital platforms. By combining text and image analysis, these models can identify potential risks more accurately, such as detecting sexual or harmful content across both written and visual formats. This multimodal approach emphasizes the importance of using different data types to broaden the scope of harmful content detection, offering a more comprehensive and effective safeguard for minors in online spaces.

In general, these studies collectively underscore the growing potential of LLMs to address the challenges of harmful communication in digital spaces, particularly in contexts involving minors. By fine-tuning models on specialized datasets, incorporating multimodal reasoning, and addressing ethical concerns, researchers are laying a foundation for safer and more responsible AI applications in safeguarding vulnerable populations [6,7].

3. Proposed Approach

Current LLMs face challenges in accurately determining and explaining whether a text contains age-specific language and/or sexual (abuse) references. To address this, we focused on fine-tuning a single model to specialize in these tasks more effectively. Using LLaMAntino as our baseline, we examined its ability to provide detailed explanations for requests like: "Analyses the following text and indicates whether the language used belongs to a minor or an adult, explaining:" and "Analyses the following text and indicates whether the language used contains sexual references or porn, explaining:".

The initial responses from the model were not sufficiently detailed, leading us to perform fine-tuning using a carefully constructed dataset. We annotated 500 synthetic samples to train and test the model, ensuring the dataset covered a variety of cases involving minor/adult language distinctions and adult sexual references. These synthetic samples were generated using ChatGPT-4o, providing to the model a more comprehensive understanding of the nuances involved in identifying and explaining these specific linguistic patterns. The fine-tuned model

aimed to enhance its accuracy and ability to provide clearer, context-sensitive explanations for such complex tasks.

3.1. Description of the solution and dataset

The initial phase of our work focused on generating synthetic phrases for two distinct tasks: identifying age-specific language and detecting sexual content in text.

For the age detection task, we defined the following parameters to categorize the content:

- *Age group*: This category helps to classify text based on the perceived age of the speaker.
 - 13 years or under
 - 14 -17 years
 - 18 years or over
- *Sentence complexity*: This parameter captures the complexity of the language used, which can vary across age groups.
 - very simple
 - simple
 - intermediate
- *Use of slang*: Identifying whether slang is present helps in determining the likelihood of the text belonging to a younger or older audience.
 - Yes
 - No

For the sexual content detection task, the following parameters were established:

- *Sexual allusion*: This identifies whether the text contains indirect or direct references to sexual matters.
 - Yes
 - No
- *Presence of emojis*: Emojis can often indicate a more casual or suggestive tone, influencing the interpretation of the text.
 - Yes
 - No
- *Reference to digital content*: This category captures whether the text mentions digital content, such as images or videos, which may relate to explicit or suggestive material.
 - None
 - Image
 - Video
- *Harmful behavior*: This parameter focuses on detecting negative attitudes such as harassment, manipulation, coercion, or objectification within the text.
 - Harassment
 - Manipulation
 - Coercion
 - Objectification
 - None

- *Explicit language*: This measures the level of explicit language used in the text, from “none” to “high”.
 - No
 - Low
 - Medium
 - High

The next step involved two datasets, one for the age detection task and one for the sexual content detection task, creating a unified collection of synthetic phrases. By generating a total of 500 samples across both tasks, we ensured a diverse representation of content. These samples were carefully annotated based on the parameters outlined above, which enabled us to fine-tune the model to better identify and explain age-specific language and sexual references in text. Subsequently, an accurate explanation for each phrase in the dataset was generated (using ChatGPT-4o).

Moreover, the explanation of each phrase, along with the previously identified categories, were grouped into a single paragraph. The final dataset, created in JSON format, was divided into a training set (450 examples) and a test set (50 examples) containing:

- *Instruction*: LLM commands
 - "Analyse the following text and indicate whether the language used belongs to a minor or an adult, explaining:"
 - "Analyse the following text and indicates whether the language used contains sexual references or porn, explaining:"
- *Input*: dataset phrases
- *Output*: explanation for each phrase

An example for each prompt is reported below:

- Age task
 - "*Instruction*": "Analyses the following text and indicates whether the language used belongs to a minor or an adult, explaining:"
 - "*Input*": " My dad took away my PlayStation because I took a punishment."
 - "*Output*": "This sentence reflects the language of children aged 13 years or younger, as the parental reaction and focus on the PlayStation is typical of this age group. In particular, it is a sentence characterised by an intermediate complexity that does not make use of slang."
- Sexual task
 - "*Instruction*": " Analyses the following text and indicates whether the language used contains sexual references or porn, explaining:"
 - "*Input*": "I imagine you in your room right now, make a video for me."
 - "*Output*": "The text presents references to digital content, in particular videos. The imagining of the person in their room, accompanied by the request for a video, evokes a private and sensual image, typical of an erotic context. It is

possible to define the attitude in question as harassment. In addition, one can assess the language used as highly suggestive of sexual themes.”

Finally, fine-tuning was performed on LLaMAntino (version 3-ANITA-8B-Inst-DPO-ITA) using the previously created dataset.

3.2. Main technical details

For the training, we chose LLaMAntino as the pre-trained model. LLaMAntino [8] was selected because it is a model designed for the Italian language and is recognized as one of the best LLMs available for this language. Since our dataset is entirely in Italian, it was the most natural choice for this experiment.

Fine-tuning of the model was performed on Kaggle using the following parameters:

- num train epochs: 10
- learning rate: $2e-4$,
- logging steps: 50,
- optimizer: adamw 8bit,
- weight decay: 0.01

We obtained the following results (1 epoch corresponds to 56 steps):



Figure 1 - Training Loss for LLaMAntino

4. Evaluation

Model evaluation was simultaneously performed for both age and sexual tasks.

Evaluation of the age task consisted of verifying how accurately the model recognized the types “Age Group”, “Sentence complexity” and “Use of Slang” by comparing them with our ground truth.

Evaluation of the sexual task consisted of verifying how accurately the model recognized the types “*Sexual allusion*”, “*Presence of emojis*”, “*Reference to digital content*”, “*Harmful behavior*” and “*Explicit language*” by comparing them with our ground truth. The accuracy metrics were assessed using a test set containing 50 sentences. We decided to give 1 point if the category type matched our ground truth and 0 points if there was no match. Then we aggregated all data and we measured partial and total accuracy.

4.1. Test results for age task

Category	Ground truth	Output	Accuracy
Age Group	25	24	96,00%
Sentence complexity	25	18	72,00%
Use of Slang	25	22	88,00%
Total	75	64	85,33%

4.2. Test results for sexual task

Category	Ground truth	Output	Accuracy
Sexual allusion	25	25	100,00%
Presence of emojis	25	25	100,00%
Reference to digital content	25	24	96,00%
Harmful behavior	25	17	68,00%
Explicit language	25	21	84,00%
Total	125	112	89,60%

4.3. Comparison among baseline and fine-tuned model

To show the differences between the fine-tuned model and the baseline (LLaMAntino version 3-ANITA-8B-Inst-DPO-ITA), we executed the two following prompts (respectively for age and sexual task), using two sentences from the test set as an example:

- Analyse the following text and indicate whether the language used belongs to a minor or an adult, explaining:
“When will I be able to go out alone without them annoying me?”

Model	Output
Baseline	I believe that the language used in this text belongs to a minor (probably a child around 4-6 years of age) in the care of a parent or an adult, as the text expresses a concern related to personal protection, typical of an age when dependence on parents is still great and awareness of one's own safety is in formation.

LLaMAntino fine-tuned	The sentence belongs to 14-17 year olds, who seek autonomy and want to show that they are capable without help by displaying a more autonomous tone. In particular, it is a sentence characterised by a simple complexity that does not make use of slang.
-----------------------	--

- Analyse the following text and indicates whether the language used contains sexual references or porn, explaining:
 “I'd like to see you without those clothes on... 😏”

Model	Output
Baseline	<p>I do not consider that the language used in the text contains sexual or pornographic references.</p> <p>The text expresses a friendly or romantic expression, with a light and non-offensive tone, as the sentence is addressed to someone known and not to an anonymous subject. The presence of an emoticon such as 😏 suggests a playful and non-sexually explicit tone. In general, the text is more akin to a polite expression and not erotic or offensive content.</p>
LLaMAntino fine-tuned	<p>The phrase emphasises the desire to see the person without clothes, implying an explicit sexual intention. The tone is made more provocative by the emoji 😏. It is possible to define the attitude in question as objectification. In addition, one can assess the language used as highly racy on issues of a sexual nature.</p>

4.4. Test the fine-tuned model on a simulated chat

To test our model in a typical use case, we simulated a chat between two people, one adult and the other a minor, and applied prompts for age recognition and sexual language detection in sequence.

- Analyse the following text and indicate whether the language used belongs to a minor or an adult, explaining:
 “User 1: You look too sexy in that photo, but I want to see you without anything. User 2: I don't know... mum says you don't do these things”

Model	Output
LLaMAntino fine-tuned	The User 1 sentence reflects sexually explicit language typical of an adult context, while the User 2 reflection introduces an ethical concern characteristic of 14-17 year old adolescents. In particular, it is a sentence characterised by an intermediate complexity that does not make use of slang.

- Analyse the following text and indicates whether the language used contains sexual references or porn, explaining:
“User 1: You look too sexy in that photo, but I want to see you without anything. User 2: I don't know... mum says you don't do these things”

Model	Output
LLaMAntino fine-tuned	The text contains references to digital content, in particular photos. The request for nudity emphasises a specific interest in explicitly sexual content. The language used is directly related to a context of an erotic nature. It is possible to define the attitude in question as coercion. In addition, one can assess the language used as highly racy on topics of a sexual nature.

5. Conclusion and limitations

As we can notice from the previously provided examples, it is evident that the baseline model generates responses that are more generic, imprecise, and less aligned with our specific goals. In contrast, the fine-tuned model produces more technical and detailed responses, indicating that the fine-tuning process has enhanced its ability to perform the proposed task effectively.

Looking ahead, there is still room for improvement. The absence of real-world data has limited the model overall accuracy, as it struggles to capture the nuances of diverse age groups and variations in sexual language. Collaboration with field experts, such as psychologists and sociologists, could help refine the selection of dataset categories and provide more precise advice and explanations tailored to the task.

Additionally, experimenting with more advanced models, such as LLaMAntino with 13B parameters, could further enhance performance. However, it is important to consider the trade-off between effectiveness and efficiency, as overly complex models may improve accuracy at the expense of practical usability.

Acknowledgments

This work was partially supported by the following projects: SERICS - “Security and Rights In the CyberSpace - SERICS” (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] Livingstone, S., & Smith, P. K. (2014). Annual research review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. *Journal of child psychology and psychiatry, and allied disciplines*, 55(6), 635–654. <https://doi.org/10.1111/jcpp.12197>
- [2] Nguyen, T. T., Wilson, C., & Dalins, J. (2023). Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.
- [3] Cook, D., Zilka, M., DeSandre, H., Giles, S., Weller, A., & Maskell, S. (2022). Can We Automate the Analysis of Online Child Sexual Exploitation Discourse?. *arXiv preprint arXiv:2209.12320*.
- [4] Lin, H., Luo, Z., Ma, J., & Chen, L. (2023). Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*.
- [5] Hu, C., Liu, B., Yin, M., Zhou, Y., & Li, X. (2024). Multimodal chain-of-thought reasoning via chatgpt to protect children from age-inappropriate apps. *arXiv preprint arXiv:2407.06309*.
- [6] Barletta, V. S., Calvano, M., & Sciacovelli, A. (2024). *Cyber social security in multi-domain operations*. In *Proceedings of the 2024 IEEE International Workshop on Technologies for Defense and Security (TechDefense)* (pp. 41–46). IEEE. <https://doi.org/10.1109/TechDefense63521.2024.10863352>
- [7] Barletta, V. S., Caivano, D., Catalano, C., de Gemmis, M., & Impedovo, D. (2024). Cyber social security education. In *Extended Reality: International Conference, XR Salento 2024, Lecce, Italy, September 4–7, 2024, Proceedings, Part IV* (pp. 240–248). Springer. https://doi.org/10.1007/978-3-031-71713-0_16
- [8] “LLaMAntino on HuggingFace”
Link: <https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA>