Shall we always use LLM Watermarking?

Alberto Vitto
University of Modena and
Reggio Emilia

Federico Betti
University of Trento

Chenyu Zhang
University of Trento

Megi Dervishi
University of Paris-Dauphine

Tural Mammadov
CISPA Helmholtz Center for
Information Security

Emmanouil Angelis
Helmholtz Munich

1 Introduction

The widespread adoption of large language models (LLMs) like ChatGPT has significantly simplified the generation of sentences, paragraphs, executable code, and images. These generated resources exhibit remarkable human-like behavior, thus exerting a substantial impact on society [1]. However, this convenience also raises concerns regarding the ability to distinguish between content generated by humans and LLMs, a crucial endeavor to mitigate the misuse of such technology for purposes like disseminating propaganda and fake news, which can harm society.

To address this challenge, [2] have proposed the use of watermarks as a potential solution. These watermarks are signals embedded within generated sentences, imperceptible to humans but detectable through algorithms. Their method offers several advantages: watermark detection does not require access to the language model API or knowledge of model parameters; watermark generation only involves inference, eliminating the need for retraining the language model; the watermark remains detectable even with a partial segment of the sentence; and it is resistant to removal without extensive modification of the generated tokens. Furthermore, statistical analysis has demonstrated the effectiveness of watermark detection.

Building upon their prior research, we utilize the Llama 2 model [3] for text generation, incorporating watermarked text to assess the capacity of both human evaluators and machine learning algorithms to differentiate between watermarked and non-watermarked content. This approach serves as a means to evaluate whether the usage of the watermarking technique impacts the final generation quality. Notably, our findings reveal that neither humans nor machine learning algorithms can effectively discern watermarked text from non-watermarked text, indicating a relatively high level of quality preservation, robustness and security in the watermarking method.

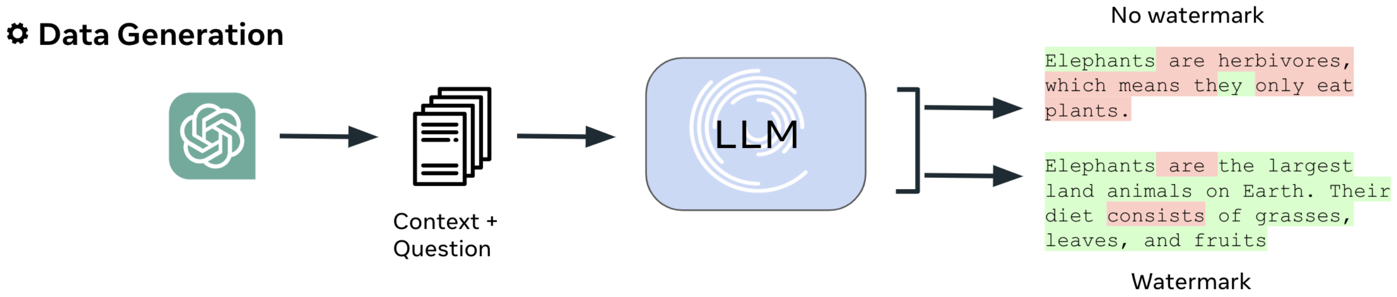
2 Methodology

The primary objective of this study was to test and compare Large Language Models during free generation and when applying the watermarking method discussed in [2]. The Llama 13B Chat model was selected for text generation due to its balanced trade-off between execution time and performance. An ideal watermarking technique should meet two primary criteria: it should not diminish the model's generative capabilities, thereby maintaining the quality of the output, while being easily identifiable and verifiable by a detection algorithm.

2.1 Manual Evaluation

The aim of the manual evaluation was to test whether a human evaluator could discriminate between text with and without a watermark. The syntactic and semantic quality of the generated text was also assessed to determine if watermarking introduced any adverse effects on the quality of generation. To facilitate this, a dataset comprising 30 paragraphs was created, each paired with a corresponding question for the model to address. This method was favored over free-text generation as it confines the content to pertinent responses. ChatGPT was employed for the generation of both the text paragraphs and the questions. The comprehensive process is depicted in Figure 1, while the questions for the human evaluator are showcased in Figure 2.

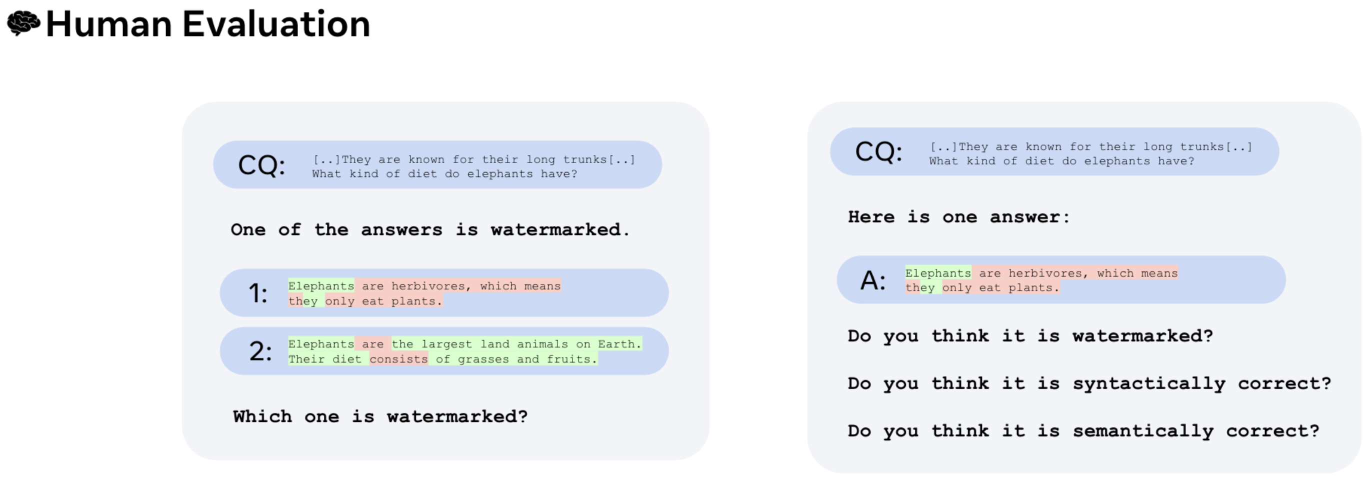
In a subsequent phase, adjustments were made to the gamma generation parameter, which led to a reduction in the percentage of words within the model's potential vocabulary. Users were then prompted to replicate the measurements and comparisons.



Elephants are herbivores, which means they only eat plants.

Elephants are the largest land animals on Earth. Their diet consists of grasses, leaves, and fruits

Figure 1: Data Generation process for Human Evaluation



[..]They are known for their long trunks[..] What kind of diet do elephants have?

One of the answers is watermarked.

1: Elephants are herbivores, which means they only eat plants.

2: Elephants are the largest land animals on Earth. Their diet consists of grasses and fruits.

Which one is watermarked?

[..]They are known for their long trunks[..] What kind of diet do elephants have?

Here is one answer:

A: Elephants are herbivores, which means they only eat plants.

Do you think it is watermarked?

Do you think it is syntactically correct?

Do you think it is semantically correct?

Figure 2: Human Evaluation Strategy

2.2 Automatic Evaluation

Beyond manual evaluation, an attempt was made to assess whether a machine learning model could differentiate between text generated with and without watermarking. This assessment started with a dataset named Alpaca [4], consisting of 52,000 texts crafted from reasoning questions and instructions. Answers were generated for each question using the Llama 13B model, both with and without watermarking. Following this, five different models were trained on a binary classification task to detect the presence of watermarking. The aim was to see if the machine learning model could identify the watermark and whether the quality of the generated text was influenced by its use. Five classification models were employed, including basic ML models like SVM and Decision Trees, as well as a Transformer model — specifically BERT [5] — for embedding extraction, complemented by a linear classification layer.

It is important to note that based on the algorithm presented in [2], there's a capability to algorithmically determine the extent of watermarking usage in the text generated using the aforementioned approach.

2.3 Implementation Details

The code for implementing watermarking and for generating watermarked text draws heavily from the approach detailed in [2]. For the experiments, the Llama 13B model was utilized. This model consumes almost the full memory capacity of a single GPU on the HPC Leonardo system, namely 56GB of memory. Some tests were also conducted with the Llama 70B model, distributed across 4 GPUs of 64 GB each (in a single node), using fp16. That was a unique chance to take advantage of the high compute resources offered in the Leonardo supercomputer, since just loading the 70B parameters model consumes 135 GB of GPU memory; the model was evenly distributed among the 4 GPUs, meaning that in each GPU 30 GBs of the model were loaded. Due to time constraints and efficiency considerations, data for both automatic and manual evaluations was generated using the 13B model.

3 Results

3.1 Manual Evaluation

Manual evaluation was conducted by the authors. When a standard watermarking approach was adopted, evaluators could not distinguish between text with and without watermarking. Conversely, when a more prominent watermark was applied, the text containing the watermark was identified more successfully. The primary differences observed were attributed to the reduced vocabulary available for watermarking. Occasionally, it was noted that the model used inappropriate words in certain contexts and often generated lengthier text as it could not directly address the discussion point. Results are shown in Figure 3 and Figure 4

Nevertheless, the grammatical and syntactical quality of models, both with and without the use of watermarking, appeared comparable, as shown in Figure 5.

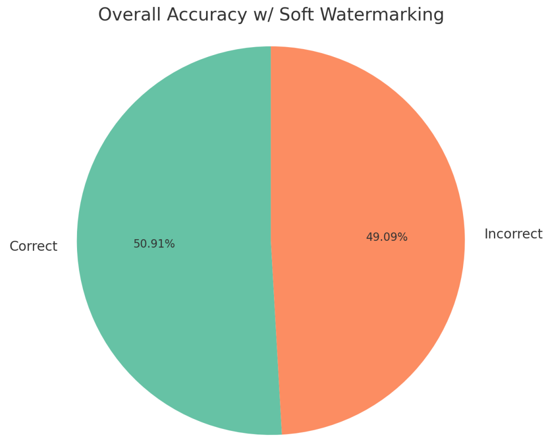


Figure 3: Manual evaluation accuracy with standard watermarking.

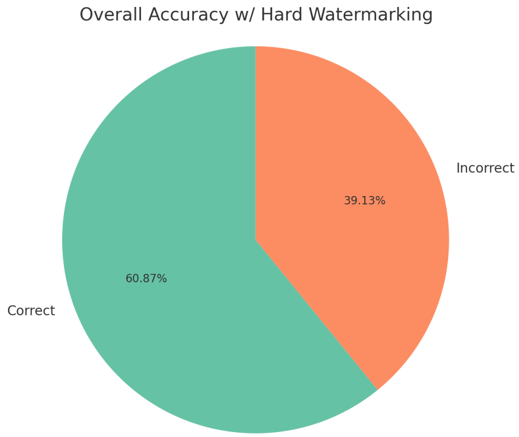


Figure 4: Manual evaluation accuracy with hard watermarking.

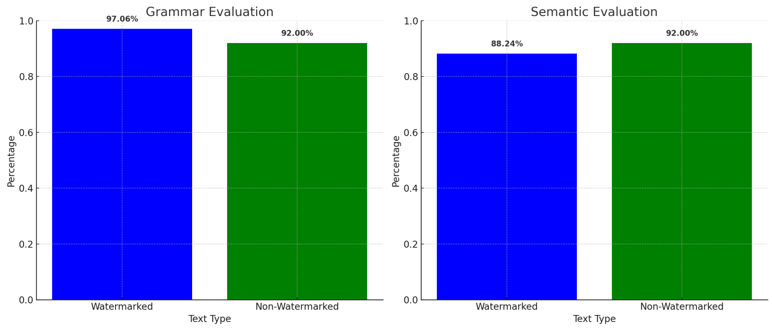


Figure 5: Comparison on Syntactic and Semantic evaluations with and without watermarking.

Classifier	Accuracy
Decision Tree	0.5010
Multi-layer Perceptron	0.5067
SVM	0.5324
SGD	0.5070
BERT + Linear	0.5247

Figure 6: Watermarking Classification accuracy

3.2 Automatic Evaluation

Automatic evaluation was carried out using only the standard watermarking application parameters. None of the models were able to learn patterns within the watermarked text, and all classifications hovered around 50% as shown in Figure 6. In a binary classification context, this indicates that the models were unsuccessful in distinguishing between the two classes.

4 Conclusion

As the use of large language models for text generation becomes increasingly prevalent in the future, addressing concerns such as data integrity, authenticity verification, and potential misuse becomes crucial. There will be a growing need for techniques capable of discerning whether the generated text originates from an automated machine. Through this study, despite the authors are aware of the clear limitations in terms of sample size, it has been highlighted that the judicious introduction of watermarking still allows for high-quality generation, while keeping the ability for a machine to detect algorithmically its the presence.

References

[1] Malak Abdullah, Alia Madain, and Yaser Jararweh. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. IEEE, 2022.

[2] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[4] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.