https://doi.org/10.2352/EI.2025.37.4.MWSF-298

Natural Language Watermarking with ChatGPT

Martin Steinebach, Raphael Antonius Frick; Fraunhofer SIT | ATHENE: Darmstadt, Germany

Abstract

Digital watermarks for texts come in numerous forms. The text itself, but also its appearance, i.e. font, letter spacing or line spacing, can be modified. Here, we present an approach that marks the text itself by introducing changes to the written words. For this, numerous methods are known, such as change from active to passive, modulation of sentence lengths or replacements with synonyms. We use ChatGPT to supplement existing texts with suggestions for synonymous formulations. We also look at evaluating the transparency of the marked texts with the help of ChatGPT.

Motivation

Natural language watermarking, a technique used to embed hidden information in text, can significantly benefit from large language models (LLMs) such as ChatGPT due to their advanced understanding of linguistic patterns and contextual nuances [8]. LLMs offer enhanced capabilities to generate semantically coherent and contextually appropriate watermarked text, reducing the risk of detection by human readers or automated systems. By leveraging their ability to produce high-quality, diverse, and natural-sounding speech, LLMs can embed watermarks in a more subtle and less intrusive manner, preserving the readability and fluency of the original text. In addition, LLMs can dynamically adjust watermarking strategies based on different linguistic contexts, ensuring robustness and adaptability across different languages, dialects and writing styles. This makes them an invaluable tool for improving the effectiveness, invisibility, and resilience of natural language watermarking.

The goal of this work is to develop a natural language watermarking system that embeds a watermark message into pre-existing text (cover), instead of generating new text as in steganography [11]. This is done by changing the words of the text, not a visual representation or modulation of whitespace or similar formatting elements. Our goal is to minimize any perceptible impact on linguistic quality and readability, and to provide a high level of transparency in the watermarking process. The watermarking function is based on finding synonyms for selected words in the text. The LLM selects words in the cover for which synonyms can be provided and also evaluates the quality of the replacement. This is done by estimating the impact of the change caused by replacing a word with its synonym.

Unlike steganography, which prioritizes concealment, this approach aims to strike a balance between maximizing embedding rate and preserving the original semantic, syntactic, and stylistic integrity of the text.

It should be noted that this work does not address the issue of watermarking output text produced by an LLM as described in the context of “labeling” generative AI content, such as in [7]. This process would label all content to make it identifiable as being generated by an LLM. Instead, we assume that the original text is written by a human and the LLM helps to embed an individual watermark to distinguish copies of the text. This concept is called transaction watermarking [10] or forensic watermarking [6] and is a common alternative to cryptography-based DRM.

State of the Art

The paper is not novel in the sense of the watermarking strategy. Papers like [13, 2] proposed NLP-based approaches 20 years ago. The novel aspect in our work is to show how well generic NLP systems like ChatGPT can be applied to execute such strategies. In previous works, highly specialized NLP systems needed to be designed often with many limitations. Today, a generic tool like ChatGPT is sufficient to achieve comparable results with an appropriate prompt. It can also be applied to verify the transparency of the suggested options by evaluating the impact of the changes to the text quality.

In [1], the Adversarial Watermarking Transformer (AWT) is introduced by Abdelnabi and Fritz as a method for embedding data within natural language text without requiring paired training data or rule-based encoding systems. AWT employs a transformer-based encoder-decoder architecture, similar to sequence-to-sequence models used in machine translation. This architecture functions as a hiding network, accepting an input sentence and a binary message to produce a modified output text. A separate transformer encoder, acting as a decoder, reconstructs the binary message from the output text. Adversarial training is utilized, wherein the hiding and decoding components are trained in opposition to an adversarial classifier designed to distinguish between the original and modified texts. The model is optimized to minimize text alterations, ensure accurate message decoding, and deceive the adversarial classifier simultaneously.

Zhang et al. in [15] discuss REMARK-LLM, a robust and efficient framework for watermarking text generated by LLMs. The framework integrates message encoding, re-parameterization, and decoding modules, which are trained jointly to embed watermark signatures into LLM-generated text while preserving semantic coherence. The encoding module embeds watermarks, and the decoding module extracts the embedded messages from watermarked text. As text generation and watermarking are combined, this approach cannot be used for forensic watermarking of existing text.

Own Previous Work

In our previous work on natural language watermarking [4] we took advantage of natural language processing methods but did not employ large language models. We introduced several methods as alternatives to the common synonym replacement strategies.

Enumeration Modulation (EM) relies on grammatical rules for reordering constituents within enumerations. For example, the country names in “In Romania, Bulgaria, and Hungary” can be reordered into other permutations, such as “In Hungary, Bulgaria, and Romania.” To address challenges, idiomatic expressions and sensitive contexts are avoided using lists of idioms and POS-tag patterns. Additionally, a classifier is employed to improve the identification of suitable patterns.

Conjunction Modulation (CM) focuses on reordering two elements joined by a conjunction, such as switching “Viktor and Anna” to “Anna and Viktor.” Challenges arise in protecting well-known fixed phrases, such as “Bonnie and Clyde,” which are safeguarded using a blacklist.

Prefix Expansion (PE) involves expanding negated words with prefixes like “un-” into explicit negations, as in transforming “unimportant” into “not important.” To avoid grammatically incorrect transformations, particularly when the context negates or modifies the prefix meaning, classifiers trained on contextual features are used.

Compound Segmentation (CS) splits complex German compounds using optional rules for hyphenation and spacing. For instance, “Kaffeeernte” can be segmented into “Kaffee-Ernte.”

The main challenge with these methods is the low bitrate. Occurrences where they can be applied without quality degradation of the text are rare, therefore in real-world experiments only roughly one bit per written page could be embedded.

Our recent work on ChatGPT presents synthetic text generation for steganography [11]. Here we use ChatGPT to create a text with a key-dependent set of words which encode the binary sequence of the hidden message.

Approach

Our proposed approach is solely based on ChatGPT prompts generating the embedding options. It is a non-blind concept requiring the original text for watermarking detection.

For easy usage and integration, we design the system in two steps: first a version of the cover with alternative words together with a transparency estimation of the change is generated. Hereby, the transparency estimation assesses the potential degree of semantic change in the text if the alternative wording were chosen during the embedding process. An example is shown in Table 1. The actual embedding process uses this version to embed a watermark message under control of a transparency threshold. Using this strategy, creating multiple copies with individual watermarks as typical in transaction watermarking, is highly efficient: the complex generation of local text alternatives which encode the message bits only is executed once. The generation of the marked copies is basic text processing requiring only little resources. The approach is similar to the concept of a watermarking container[12] which is commonly used for audio and video content.

Embedding The embedding process consists of multiple steps. First, the alternatives are generated, afterward the transparency of the alternatives can be verified. Then the actual marked cover is generated by choosing the alternatives controlled by a watermarking message.

This is an example prompt to create the watermarking option text:

⇒ *Generate an altered version of the text “cover” by providing the original word or short phrase and one alternative for each. Each pair should be formatted in brackets [] separated by a slash /, indicating the choice. Include stop words where necessary to prevent errors, such as ”a [cow/eel]” becoming ”[a cow/an eel]”. Additionally, provide the estimated change strength directly behind each pair of alternatives, ranging from 0 (minimal change) to 99 (moderate change). Write the change strength in (). Avoid changing names and numbers. The objective is to illustrate diverse rephrasing options while ensuring clarity and coherence in the modified text.*

Changes in named entities and numbers are to be avoided as they are likely to cause major changes in the content. The estimated change strength is a first indicator of the resulting transparency if the second alternative is used to embed a bit value 1 at that position. To further verify the resulting transparency, we can use an additional prompt:

⇒ *For evaluation, create a version of cover (called m1cov) where always the second alternative (the new one) is chosen. Compare the text quality of cover and m1cov. If there are passages that produce a problem in the text quality of m1cov, set the change strength at that position on value 99.*

To generate a marked cover, we use a simple Python script parsing the cover with the alternatives and selecting words corresponding to a binary watermarking message. The transparency of the watermark can be controlled by providing a threshold of the accepted change strength. Using a script here is more efficient to generate a large set of marked covers with random watermarking messages.

Detection Apart from marking the text, the detection of the watermark is also done using a prompt and an LLM. Nonetheless, an algorithmic method can be used as well. For this, the cover text containing alternatives for certain text positions is parsed which is then compared with the marked cover to identify which alternatives were chosen and thereby which bit was embedded. The latter approach has the advantage of being more efficient, especially when scanning a large amount of text for embedded watermarks.

⇒ *You are a watermarking detector. When I provide you with a ”cover” including alternatives in brackets divided by a backslash and a ”marked cover” where one of the alternatives has been used to code a binary sequence, you provide the embedded message by checking the marked cover for each position of alternatives if the first or second option from the cover has been used and print a 0 for the first and a 1 for the second option. Ignore all positions with a change strength above the parameter ”transparency”.*

Blind vs. non-blind Watermarking

In many application scenarios blind watermarking is preferred, if not required. Blind watermarking allows retrieving an embedded message without knowledge of the cover. Some text watermarking strategies like the ones introduces by us in [4] allow blind watermarking as there exactly two options for one given text element, like using Conjunction Modulation and testing if the cryptographic hash value of the first item is bigger than that of the second item or not. A detector will find the conjunction, compute the cryptographic hashes of the items and provide the result by a simple comparison.

As the output provided by ChatGPT is not deterministic, the results obtained are too unstable to enable such a blind strategy.

Input	Output
One boy spoke up. He knew 25 x 200 = 5,000, so he guessed that 25 x 199 could equal 4,999. Baar paused before breaking the equation down into smaller chunks, offering a reminder that multiplication is just adding the same number over and over again. The group slowly caught on to the pattern until they produced the right answer: 4,975. "I know that they can do it," he said after the session. "It's a lot of giving them encouragement." At Perry Street Prep, each of its roughly 450 students received extra help in math this past year, whether in a tutoring group like the one Barr led or more time with a teacher — an intensive push to regain academic ground lost during the pandemic. Children across the country returned from virtual learning much further behind in math — a subject where each skill builds on another and gaps in understanding can leave students lost — than they were in reading. Between fall 2019 and 2022, national test scores sank four points in reading and nine points in math for 13-year-olds — the largest drop in math in half a century. And while scores have started to improve, researchers at the testing nonprofit NWEA estimated last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to catch up in math, compared with seven extra months in reading.	One [boy/student] (10) spoke up. He [knew/understood] (15) 25 x 200 = 5,000, so he [guessed/assumed] (20) that 25 x 199 could [equal/be equal to] (5) 4,999. Baar [paused/hesitated] (10) before [breaking/splitting] (10) **the equation [down/into parts] (5) into [smaller/tinier] (10) chunks**, [offering/giving] (10) a [reminder/notice] (15) that multiplication is just [adding/summing] (10) the same number [over and over again/repeatedly] (20). The group [slowly/gradually] (5) caught on to the pattern until they [produced/derived] (20) the right answer: 4,975. "I know that they can [do it/manage it] (5)," he said after the session. "It's a lot of [giving/providing] (10) them [encouragement/support] (20)." At Perry Street Prep, each of its [roughly/about] (5) 450 students [received/got] (5) extra help in math this past year, whether in a [tutoring/coaching] (10) group like the one Barr led or [more/additional] (5) time with a teacher — an [intensive/focused] (15) push to regain academic ground lost during the pandemic. Children across the country [returned/came back] (5) from virtual learning much [further/more] (10) behind in math — a subject where each skill builds on another and gaps in understanding can [leave/put] (5) students lost — than they were in reading. Between fall 2019 and 2022, national test scores [sank/fell] (5) four points in reading and nine points in math for 13-year-olds — the [largest/biggest] (5) drop in math in half a century. And while scores have started to improve, researchers at the testing nonprofit NWEA [estimated/projected] (10) last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to [catch up/recover] (10) in math, compared with seven extra months in reading.

Table 1: Input text and output (watermarking options).
Input https://www.washingtonpost.com/education/2024/07/01/math-learning-loss-pandemic-recovery

Alternatives provided as well as suitable positions in the text differ from prompt call to prompt call. As an example, we use the first sentence of MSFW introduction:

Original MSFW 2025 Overview: "The exponential growth in digital media accessibility, spanning images, audio, video, graphics, and text, has sparked the development of novel applications, while also introducing a host of significant security challenges and risks."

Result 1: The [exponential/rapid] (30) growth in digital media accessibility, spanning [images/pictures] (10), [audio/sound] (15), [video/visuals] (25), [graphics/designs] (20), and [text/written content] (10), has sparked the development of [novel/new] (15) applications, while also introducing a host of [significant/serious] (20) security [challenges/issues] (15) and risks.

Result 2: The exponential [growth/expansion] (15) in digital media [accessibility/availability] (10), spanning images, audio, video, graphics, and text, has [sparked/triggered] (20) the development of novel [applications/uses] (10), while also [introducing/posing] (25) a host of significant security [challenges/concerns] (20) and risks.

As a result, our proposed approach requires a strong non-blind strategy. Strong means that not only the original cover needs to be available, but also the alternatives suggested by ChatGPT. Otherwise, one could only compare the cover and the marked cover and find all positions where words have been changed. In this case, only sequences of 1s could be retrieved because we only identify the positions where alternative 2 was used.

General Watermarking Design Options

The focus of this work is the use of ChatGPT to provide text alternatives for natural language watermarking. However, some general aspects should be briefly discussed. The security of the watermark against attacks is limited. Using ChatGPT with the embedding prompt on the marked cover is likely to produce collisions of alternative positions and thus allow distortion of the embedded message. The robustness of this type of text watermark is of limited relevance. Noise and quality degradation as in audio or video due to scaling and lossy compression will not occur. The most comparable process might be printing, scanning, and applying OCR. This could introduce noise in the form of OCR errors. But modifying the watermark detection to be robust against these errors should be straightforward, since the only challenge is to decide which alternative was used.

This method does not use a traditional watermarking key that controls the position of the watermark or similar aspects. The "key" is the cover with the alternatives created in the first step of the process. Of course, a key can be used to encrypt the embedded message.

Evaluation

As a first proof of concept, we use the embedding prompt on a single news article and discuss the results. We show an example of a source (cover) and the watermarking options in Table 1. Identified watermarking options are listed in Table 2. Two resulting example marked texts are given in Table 3.

One mistake made by the generation of embedding options is shown in bold case in the output from Table 1. The section "... equation [down/into parts] (5) into [smaller/tinier] (10) chunks..." can produce the word sequence "...equation into parts into smaller chunks...". This causes confusion during reading. One simple countermeasure would be to enforce a minimal distance of embedding positions to prevent interaction of watermarking options.

Detection of the embedded message using the reference (example shown in the Table 5) is, as expected, reliable and straightforward: the marked cover is scanned for positions of alternatives, the choice is noted, and the corresponding bit is added to the sequence.

No.	Term	Options
1	student	boy / student
2	understood	knew / understood
3	guessed	guessed / assumed
4	equal	equal / be equal to
5	paused	paused / hesitated
6	breaking	breaking / splitting
7	down	down / into parts
8	smaller	smaller / tinier
9	giving	offering / giving
10	notice	reminder / notice
11	adding	adding / summing
12	repeatedly	over and over again / repeatedly
13	slowly	slowly / gradually
14	manage it	do it / manage it
15	providing	giving / providing
16	encouragement	encouragement / support
17	about	roughly / about
18	got	received / got
19	tutoring	tutoring / coaching
20	more	more / additional
21	focused	intensive / focused
22	came back	returned / came back
23	more	further / more
24	fell	sank / fell
25	largest	largest / biggest
26	projected	estimated / projected
27	recover	catch up / recover

Table 2: Term and Replacement Table

We ran a limited number of tests to verify that the approach was generic. As performance indicators, we calculated the number of suggested alternatives divided by the number of words in the passage (BpW = Bits per Word). We also calculated the expected number of bits to be embedded per page (BpP). A page was estimated to have 275 words. The covers were taken from Project Gutenberg[1] and the NBC News website[2]. The first seven examples in Table 4 are from the former, the last three from the latter. News articles seem to produce fewer suggested alternatives. For the news articles, the payload would not be sufficient to encode a user ID, for example, to monitor content theft, since we can expect only a few pages per article. For Project Gutenberg content, the payload would be sufficient to embed user IDs with typical watermark lengths of 32 to 64 bits, often even on a single page.

Standard readability measures [3] like Flesch Kincaid Reading Ease are not influenced by the marking process. This is not surprising as they are based on word and syllables counts which are not significantly changed by the embedding process. Flesch Kincaid Reading Ease scores for the original is 64.2, for the two marked versions scores were 65.2 and 62.8. It must be noted that while these measures are widely used, there is also criticism about their actual meaningfulness [9].

We also used ChatGPT 4o to compare the original and a marked example: "Text 1 (the original) remains the higher-quality text in terms of readability, linguistic accuracy, coherence, and style. Text 2 (the second marked version) is coherent and understandable but contains minor linguistic inaccuracies and stylistic redundancies that make it weaker than Text 1.". The LLM especially mentions "into parts into tinier chunks" is redundant, as "parts" and "chunks" mean the same thing."

A reliable evaluation of the resulting text quality would require a standardized procedure with human readers.

Table 7 shows the result of the quality score prompt above, which identifies some problematic areas that could occur in the marked cover. Table 8 shows the alteration in change strength assigned to the word pairs. Based on this, transparency can be improved by the cost of capacity.

Watermark detection is not blind, as can be seen in the prompt above. An example of detection results for the marked cover on the left from Table 3 is given in Table 5.

Table 6 shows the reason this approach is a non-blind watermarking strategy, requiring the cover with a set of alternatives for detecting the watermark. Proprietary LLMs, like ChatGPT, do not use seed values like GenAI image generators do. The results of a task show a certain randomness, so to detect the watermark reliably, one needs the full reference. To address this issue, open-source LLMs such as LLama [14] and Mistral [5] can serve as alternatives as they provide options to enforce deterministic text generation.

Discussion

The experiments conducted in this work demonstrate that the proposed concept is functional and effectively supports the desired outcomes. The results indicate that the embedding method can successfully embed information in cover texts without significant loss of quality or transparency, thus validating the feasibility of the approach.

The transparency of the tagged cover text was found to be satisfactory, especially when appropriate precautions are taken to avoid errors during the embedding process. By enforcing rigorous cross-checking procedures, the likelihood of errors is minimized,

[1] www.gutenberg.org/ebooks

[2] www.nbcnews.com

01010101000001110100001010000	01000011101111000101011101100
One boy spoke up. He understood 25 x 200 = 5,000, so he guessed that 25 x 199 could be equal to 4,999. Baar paused before splitting the equation down into tinier chunks, offering a reminder that multiplication is just adding the same number over and over again. The group slowly caught on to the pattern until they derived the right answer: 4,975. “I know that they can manage it,” he said after the session. “It’s a lot of providing them encouragement.” At Perry Street Prep, each of its about 450 students received extra help in math this past year, whether in a tutoring group like the one Barr led or more time with a teacher — an intensive push to regain academic ground lost during the pandemic. Children across the country came back from virtual learning much further behind in math — a subject where each skill builds on another and gaps in understanding can put students lost — than they were in reading. Between fall 2019 and 2022, national test scores sank four points in reading and nine points in math for 13-year-olds — the largest drop in math in half a century. And while scores have started to improve, researchers at the testing nonprofit NWEA estimated last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to catch up in math, compared with seven extra months in reading.	One boy spoke up. He understood 25 × 200 = 5,000, so he guessed that 25 × 199 could equal 4,999. Baar paused before breaking the equation into parts into tinier chunks, giving a reminder that multiplication is just summing the same number repeatedly. The group gradually caught on to the pattern until they derived the right answer: 4,975. “I know that they can do it,” he said after the session. “It’s a lot of giving them encouragement.” At Perry Street Prep, each of its about 450 students received extra help in math this past year, whether in a coaching group like the one Barr led or more time with a teacher — an focused push to regain academic ground lost during the pandemic. Children across the country came back from virtual learning much more behind in math — a subject where each skill builds on another and gaps in understanding can leave students lost — than they were in reading. Between fall 2019 and 2022, national test scores fell four points in reading and nine points in math for 13-year-olds — the biggest drop in math in half a century. And while scores have started to improve, researchers at the testing nonprofit NWEA estimated last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to catch up in math, compared with seven extra months in reading.

Table 3: Two marked covers compared. Marked by 29 bit random message stated at the top.

Title and Author	Words	Alt	BpW	BpP
A History of Art for Beginners and Students: Painting, Sculpture, Architecture *Clara Erskine Clement Water*	290	18	0.06	19
Bad and Mad *W. C. Tuttle*	429	44	0.10	69
Larry Dexter and the Bank Mystery *Raymond Sperry*	248	10	0.04	9
Elementary Cryptanalysis *Helen Gaines*	254	26	0.10	24
Frankenstein; Or, The Modern Prometheus *Mary Wollstonecraft Shelley*	384	44	0.11	61
From Tenderfoot to Golden Eaglet *Amy Ella Blanchard*	263	18	0.07	17
Monograms & Ciphers *A. A. Turbayne*	430	18	0.04	28
Unburned Areas Are Really a Concern *Marlene Lenthang*	137	9	0.07	4
Extremely Dangerous’ Fire Conditions *Patrick Smith*	121	7	0.06	3
JD Vance Says Violent Jan. 6 Rioters Shouldn’t Receive Pardons *Alexandra Marquez and Alex Tabet*	391	5	0.01	7

Table 4: Ten examples for covers and their performance. BpW = Bit per Word, BpP= Bit per Page

ensuring text quality while still embedding the required information.

The payload capacity of the method has been shown to be sufficient for specific applications such as transactional watermarking in e-books. The amount of information that can be embedded in the cover text is sufficient for watermarking purposes, where small but critical information needs to be discreetly inserted into the content without overwhelming the natural flow of the text.

While the current experiments provide a solid foundation for the proposed concept, they are inherently limited in scope. The variety of cover texts tested and the range of potential use cases remain constrained within the boundaries of this initial study. Expanding the experiments to include a broader set of cover texts from different domains and extending the evaluation to real-world applications will be essential to assessing the robustness and applicability of the method.

The paper is not novel in the sense of the watermarking strategy. Papers like [13, 2] proposed NLP-based approaches 20 years ago. The novel aspect in our work is to show how well generic NLP systems like ChatGPT can be applied to execute such strategies. In previous works highly specialized NLP systems needed to be designed, often with many limitations. Today a generic tool like ChatGPT is sufficient to achieve comparable results with the right prompt. It can also be applied to verify the transparency of the suggested options by evaluating the impact of the changes to the text quality.

Future Work

Future work should involve the full evaluation of the proposed method using a large and diverse set of cover texts across various categories. This will ensure that the approach’s performance is robust and adaptable to different types of textual data, thus assessing its generalizability and scalability.

A key area for further investigation is the creation of a standardized methodology for evaluating the quality of the marked cover texts, particularly focusing on transparency assurance. This would involve assessing the degree to which the embedded watermarks do not affect the readability or meaning of the cover text, and ensuring that transparency remains intact while embedding

Alternatives	Chosen Option	Bit
[boy/student]	boy	0
[knew/understood]	understood	1
[guessed/assumed]	guessed	0
[equal/be equal to]	be equal to	1
[paused/hesitated]	paused	0
[breaking/splitting]	splitting	1
[down/into parts]	down	0
[smaller/tinier]	tinier	1
[offering/giving]	offering	0
[reminder/notice]	reminder	0
[adding/summing]	adding	0
[over and over again/repeatedly]	over and over again	0
[slowly/gradually]	slowly	0
[produced/derived]	derived	1
[do it/manage it]	manage it	1
[giving/providing]	providing	1
[encouragement/support]	encouragement	0
[roughly/about]	about	1
[received/got]	received	0
[tutoring/coaching]	tutoring	0
[more/additional]	more	0
[intensive/focused]	intensive	0
[returned/came back]	came back	1
[further/more]	further	0
[leave/put]	put	1
[sank/fell]	sank	0
[largest/biggest]	largest	0
[estimated/projected]	estimated	0
[catch up/recover]	catch up	0

Table 5: Detected Binary Message

the information.

The research could also move towards blind methods, where the process of embedding and detecting the hidden information can occur without prior knowledge of the marking positions. This would include developing techniques for the stable and reliable identification of optimal marking positions, as well as designing methods for mapping bit values to suitable alternative words or phrases to ensure blind detectability.

Acknowledgments

This research work has been funded by BMBF and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.

[2] Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, pages 185–200. Springer, 2001.

[3] James C Brewer. Measuring text readability using reading level. In *Advanced methodologies and technologies in modern education delivery*, pages 93–103. IGI Global, 2019.

[4] Oren Halvani, Martin Steinebach, Patrick Wolf, and Ralf Zimmermann. Natural language watermarking for german texts. In *Proceedings of the first ACM workshop on Information hiding and multimedia security*, pages 193–202, 2013.

[5] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[6] Serap Kirbiz, Aweke N Lemma, Mehmet U Celik, and Stefan Katzenbeisser. Decode-time forensic watermarking of aac bitstreams. *IEEE Transactions on Information Forensics and Security*, 2(4):683–696, 2007.

[7] Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *arXiv preprint arXiv:2409.00089*, 2024.

[8] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 2024.

[9] Stephen L Mailloux, Mark E Johnson, Dennis G Fisher, and Timothy J Pettibone. How reliable is computerized assessment of readability? *Computers in nursing*, 13:221–221, 1995.

[10] Christian Neubauer, Martin Steinebach, Frank Siebenhaar, and Joerg Pickel. Robustness evaluation of transactional audio watermarking systems. In *Security and Watermarking of Multimedia Contents V*, volume 5020, pages 12–20. SPIE, 2003.

[11] Martin Steinebach. Natural language steganography by chatgpt. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pages 1–9, 2024.

[12] Martin Steinebach, Enrico Hauer, and Patrick Wolf. Efficient watermarking strategies. In *Third International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'07)*, pages 65–71. IEEE, 2007.

[13] Mercan Topkara, Cuneyt M Taskiran, and Edward J Delp III. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 441–452. SPIE, 2005.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2:

Alternative A	Alternative B
One [boy/student] (10) spoke up. He knew [25 x 200 = 5,000/**that 25 times 200 equals 5,000**] (15), so he [guessed/**assumed**] (10) that 25 x 199 could equal [4,999/**4,999 exactly**] (5). Baar [paused/halted briefly] (20) before [breaking down/splitting] (15) the equation into smaller [chunks/sections] (5), offering a [reminder/prompt] (5) that multiplication is just [adding/repeating] (10) the same number over and over again. The group slowly [caught on/understood] (15) to the pattern until they produced the [right/correct] (5) answer: 4,975. "I know that they can do it," he said after the [session/class] (10). "It's a lot of [giving/providing] (5) them encouragement." At Perry Street Prep, each of its roughly [450/about 450] (5) students received extra help in math this past year, whether in a tutoring group like the one Barr led or more time with a teacher — an [intensive/strong] (15) push to [regain/recover] (10) academic ground lost during the pandemic. Children across the country returned from virtual learning [much further behind/significantly behind] (10) in math — a subject where each skill [builds upon/develops from] (10) another and gaps in understanding can leave students [lost/confused] (10) — than they were in reading. Between fall 2019 and 2022, national test scores [sank/dropped] (15) four points in reading and nine points in math for 13-year-olds — the largest drop in math in half a century. And while scores have started to [improve/recover] (10), researchers at the testing nonprofit NWEA [estimated/predicted] (5) last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to [catch up/make up for lost time] (15) in math, compared with seven extra months in reading.	One [boy/child] (10) spoke up. He knew [25 x 200 = 5,000/**25 times 200 is 5,000**] (10), so he [guessed/**thought**] (5) that 25 x 199 could equal [4,999/**close to 5,000**] (15). Baar [paused/hesitated] (20) before [breaking/splitting] (10) the equation down into smaller [chunks/steps] (5), offering a [reminder/note] (5) that multiplication is just [adding the same number repeatedly/repeating the same number through addition] (15). The group slowly [caught on/understood] (20) to the pattern until they produced the [right/correct] (5) answer: 4,975. "I know that they can do it," he said after the [session/lesson] (5). "It's a lot of [giving/offering] (5) them encouragement." At Perry Street Prep, each of its roughly [450/about 450] (5) students received extra help in math this past year, whether in a tutoring group like the one Barr led or more time with a teacher — an [intensive/strong] (10) push to [regain/recover] (10) academic ground lost during the pandemic. Children across the country returned from virtual learning [much further behind/significantly behind] (10) in math — a subject where each skill [builds on/develops from] (5) another and gaps in understanding can leave students [lost/confused] (10) — than they were in reading. Between fall 2019 and 2022, national test scores [sank/dropped] (10) four points in reading and nine points in math for 13-year-olds — the largest drop in math in [half a century/50 years] (20). And while scores have started to [improve/get better] (5), researchers at the testing nonprofit NWEA [estimated/predicted] (5) last summer that the average American eighth-grader would still need the equivalent of [nine more/around nine additional] (5) months of schooling to [catch up/recover the lost progress] (10) in math, compared with [seven extra months/seven more months] (5) in reading.

Table 6: ChatGPT will not provide identical alternatives if asked several times. Here are two additional examples of change options. First three different options are bold.

Open foundation and fine-tuned chat models, 2023.

[15] Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1813–1830, 2024.

Author Biography

Martin Steinebach is the manager of the Media Security and IT Forensics division at Fraunhofer SIT. From 2003 to 2007 he managed the Media Security in IT division at Fraunhofer IPSI. He studied computer science at the Technical University of Darmstadt and finished his diploma thesis on copyright protection for digital audio in 1999. In 2003 he received his PhD at the Technical University of Darmstadt for this work on digital audio watermarking. In 2016 he became honorary professor at the TU Darmstadt. He gives lectures on Multimedia Security as well as Civil Security. He is Principle Investigator at ATHENE and represents IT Forensics and AI security. Before he was Principle Investigator at CASED with the topics Multimedia Security and IT Forensics.

Raphael Antonius Frick is a research associate at Fraunhofer SIT / ATHENE Center in Darmstadt and a PhD candidate at Technische Universität Darmstadt. In the "Media Security & IT Forensics" department, he researches new detection methods for AI-generated and manipulated audiovisual data, as well as techniques for identifying disinformation in social media across multiple modalities. Additionally, he investigates the possibilities of synthetic data to achieve other security goals such as anonymity and to improve the robustness and performance of deep learning classifiers.

Problematic Area	Explanation
”be equal to” instead of ”equal”	While not a severe issue, it sounds slightly more formal and less fluid in casual contexts.
”hesitated” instead of ”paused”	”Hesitated” may imply indecision, which could alter the intended meaning.
”notice” instead of ”reminder”	”Notice” feels less natural compared to ”reminder” in this context.
”summing” instead of ”adding”	”Summing” sounds more formal and technical.
”repeatedly” instead of ”over and over again”	Both are clear, but ”repeatedly” is more technical.
”giving a notice”	Feels awkward compared to ”offering a reminder.”
”put students lost” instead of ”leave students lost”	This phrase is awkward. ”Leave students lost” is more natural.

Table 7: Problematic Areas

Original suggestions	Review results
One [boy/child] (10) spoke up. He [knew/understood] (15) 25 x 200 = 5,000, so he [guessed/assumed] (20) that 25 x 199 could [equal/be equal to] (5) 4,999. Baar [paused/hesitated] (10) before [breaking/splitting] (10) the equation [down/into parts] (5) into [smaller/tinier] (10) chunks, [offering/giving] (10) a [reminder/notice] (15) that multiplication is just [adding/summing] (10) the same number [over and over again/repeatedly] (20). The group [slowly/gradually] (5) caught on to the pattern until they [produced/derived] (20) the right answer: 4,975. “I know that they can [do it/manage it] (5),” he said after the session. “It’s a lot of [giving/providing] (10) them [encouragement/support] (20).” At Perry Street Prep, each of its [roughly/about] (5) 450 students [received/got] (5) extra help in math this past year, whether in a [tutoring/coaching] (10) group like the one Barr led or [more/additional] (5) time with a teacher — an [intensive/focused] (15) push to regain academic ground lost during the pandemic. Children across the country [returned/came back] (5) from virtual learning much [further/more] (10) behind in math — a subject where each skill builds on another and gaps in understanding can [leave/put] (5) students lost — than they were in reading. Between fall 2019 and 2022, national test scores [sank/fell] (5) four points in reading and nine points in math for 13-year-olds — the [largest/biggest] (5) drop in math in half a century. And while scores have started to improve, researchers at the testing nonprofit NWEA [estimated/projected] (10) last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to [catch up/recover] (10) in math, compared with seven extra months in reading.	One [boy/child] (10) spoke up. He [knew/understood] (15) 25 x 200 = 5,000, so he [guessed/assumed] (20) that 25 x 199 could [equal/be equal to] (99) 4,999. Baar [paused/hesitated] (99) before [breaking/splitting] (10) the equation [down/into parts] (5) into [smaller/tinier] (10) chunks, [offering/giving] (10) a [reminder/notice] (99) that multiplication is just [adding/summing] (99) the same number [over and over again/repeatedly] (99). The group [slowly/gradually] (5) caught on to the pattern until they [produced/derived] (20) the right answer: 4,975. “I know that they can [do it/manage it] (5),” he said after the session. “It’s a lot of [giving/providing] (10) them [encouragement/support] (20).” At Perry Street Prep, each of its [roughly/about] (5) 450 students [received/got] (5) extra help in math this past year, whether in a [tutoring/coaching] (10) group like the one Barr led or [more/additional] (5) time with a teacher — an [intensive/focused] (15) push to regain academic ground lost during the pandemic. Children across the country [returned/came back] (5) from virtual learning much [further/more] (10) behind in math — a subject where each skill builds on another and gaps in understanding can [leave/put] (99) students lost — than they were in reading. Between fall 2019 and 2022, national test scores [sank/fell] (5) four points in reading and nine points in math for 13-year-olds — the [largest/biggest] (5) drop in math in half a century. And while scores have started to improve, researchers at the testing nonprofit NWEA [estimated/projected] (10) last summer that the average American eighth-grader would still need the equivalent of nine more months of schooling to [catch up/recover] (10) in math, compared with seven extra months in reading.

Table 8: Based on ChatGPT review, problematic areas can be set to a high change strength of 99.