Watermarking LLM Output: Cheat Detection En Masse
ECE 570 Research Project (11/2023)

Anonymous Authors[1]

Abstract

With the rapid proliferation of ever-advancing Large-Language-Models and growing concern in academic, security, foreign policy, and political circles, historical and current methods of reliable *watermarking* are explored. Prior works are summarized and evaluated, their methods are retested for validity, and the state-of-the-art is tested both in its designed environment of Facebook OPT as well as a new environment in GPT-2. Hyperparameter adjustments are investigated across these models and their impact is compared. Real-world attacks on these security devices are also performed under recommended settings to determine the robustness of the methods.

Soft Watermarking, the primary method explored, proves a robust and oft-irrefutable indicator of academic dishonesty, fraudulent work, and massively-generated content. With p-values of orders smaller than 10^{-20} frequently appearing in naive copy-pasted text and detection evasion proving difficult for time-efficient methodology, the soft watermark appears as a very promising tool in combating malicious uses of LLMs. In addition, concerns such as large bias against non-native English speakers or nonstandard English dialects are addressed with the probabilistic approach over other Neural-Network or synonym-based methodologies.

1. Introduction

At the beginning of the Fall 2023 semester, Dr. Mark Johnson approached his ECE 337 teaching assistant body with a unique addition to the syllabus: a ChatGPT/Large Language Model clause. As a professor well versed in cheat detection, it was unsurprising he was willing to get ahead of the issue (Johnson et al., 2004). This was not the only class that had such a clause, though it was certainly the most thorough - they were becoming incredibly prevalent across all departments and course levels. It remains unclear how plausible a hard-working student slipping a fully ”GPT-ed” essay, discussion post, code snippet, or SystemVerilog module past a watchful professor is in 2023, but issues in this vein will no doubt only continue escalating in scale and frequency in the coming years.

This incident was the spark of motivation for choosing this topic in AI/ML research. That, alongside the disheartening feeling of watching my peers trivialize assignments that otherwise require care and effort without generative tools. Assignments, in an ideal educational setting, provide additional learning and topical reinforcement to students, and those benefits were stripped away by the availability of GPT-3, GPT-3.5, Bard, Llama, and now even more by the ever-improving GPT-4. Generative models are now attacking college coursework in the same manner that Sparknotes has plagued high school English classes for the last decade.

2. Background

Soft Watermarking is a technique discussed at length in this work and its related papers that seeks to leave an unmistakable trace within machine-generated text. It is embedded into a model at generation time, ideally locked behind an API in the way that GPT-4 works currently. The specific implementation seeks to adjust the output distribution of words, more formally known as ”tokens” in LLM spheres, to create an wildly improbable output that still maintains the readability and prompt accuracy of the text, also known as text ”utility”. Provided the watermark parameters and a watermark hash, a very efficient decoder algorithm can be run over the text to evaluate the likelihood for an entire text being watermarked, or a subsection of that text being watermarked.

3. Primary Paper: A Watermark for Large Language Models

3.1. Storyline

Motivation Given the recent proliferation of Transformer Models, Large-Language Models, and specifically Generative Pre-trained Transformers, or more commonly GPTs, machine text generation is more accessible than ever. Any-one with an internet connection can now get a fast, human-like response to a query without much effort. While this has enabled developments in programming efficiency with Github CoPilot, online research by exploiting search-engine style behavior, and accelerating many menial tasks, this availability also opens the door to malicious uses.

Machine Text Generation makes creating a variety of text for one purpose a trivial task. The paper detailed herein, (Kirchenbauer et al., 2023a), lists a number of problematic use cases that LLMs can enable, including:

- Machine-Generated "influencing campaigns" in political spheres
- Social Engineering, Scamming (Particularly Phishing), and Fraud
- Widespread cheating in Academia at all grade levels

There has already been an enormous push by professors to target ChatGPT and other accessible LLMs in syllabi, but one wonders how easy it would be for a professor to quickly and effectively differentiate between a machine-written essay and a human essay without perfect wording. As GPTs get better by the year, this line only blurs. It is clearly important to be able to determine machine-generated text from human-generated text, but this stands counter to the point of LLMs: to create human language. It is incredibly difficult to "mark" text as machine-generated without degrading the quality of that text and, therefore, the invisibility of that mark.

Prior Work This area of research is referred to as "steganography", the science of hiding additional information within some sort of message or data, or more commonly *watermarking*. Prior work in this area has focused more on continuous data, which is far easier to encode, or on hard rules.

The earliest of these attempts on digital text used 'synonym tables', replacing any iterations of an element in the table with a single version of the word or a small subset of possible synonyms. It was found that this significantly degraded text quality, as the connotation/denotation divide created far less human text.

Following this, more recent work has focused on vocabulary partitioning rules, which divide an LLM's vocabulary up into chunks and assign particular subsets to specific sequence indices. This has less of an impact on readability, but proved still far too disruptive to be undetectable.

Research Gap The point still remains that there is a huge need for machine-generated text watermarking for detection and inhibiting the use of LLMs for malicious purposes, but a method that can be used widespread needs to be far less intrusive for humans in order to fly under the radar.

Proposed Contributions Kirchenbauer et al. propose a watermarking encoding and detection scheme that works on-the-fly without degrading readability (using the metric of perplexity to measure). The scheme also includes privatization of watermarks through salting and hashing to create robustness against attacks to the point that any sufficient modification to text to hide a watermark is prohibitively expensive to automate or must be human-generated.

3.2. Proposed Solution

The paper contains three components:

- A Watermarking Engine embeddable into any machine learning model as an additional output layer, running on a new watermarking algorithm.
- A simple detector module which runs a statistical test on a text input given the approved and marked words.
- A key parameter which can be used to generate a secure, highly resilient watermarking tool that can be locked behind an API for generation and detection.

3.3. Claims/Evidence

Primary Claim The foremost claim of the paper is that the watermarking model produces a substantial, detectable watermark without degrading text quality.

Primary Evidence The plots in Figure 2 support that a perplexity of less than 2 can be achieved with a z-score greater than 10 for watermarked text, indicating very little impact on text quality (Perplexity of 1 is ideal), and very high certainty.

Secondary Claim Secondarily, the work claims the model is sufficiently robust to attacks attempting to "de-watermark" outputs, as substantially altering the text in an automated manner that does not change the utility or readability of the text is prohibitively costly.

Secondary Evidence Figure 5 of the work details a T5 Span attack, a replacement model for which each iteration costs roughly 1/3 of a token iteration for a LLM model. The figure demonstrates at least 30% of tokens must be replaced to achieve some sort of watermark degradation, and often more than 50% must be replaced to generate reasonable watermark detector avoidance.

Tertiary Claim The paper claims that the model is useable in other LLMs compatible with the "logits" package, which is the package used in many LLM designs prior to the softmax layer to create a layer of weighted tokens corresponding to a probability distribution.

Teritiary Evidence The linked repository contains modular code based on the logits framework that converts a standard output into one that is watermarked. This is demonstrated primarily on the Facebook OPT 6.7b, but is also capable with OPT 1.3 as discussed in the paper and is compatible with other open-source pretrained transformer architectures.

3.4. Critique/Discussion

Kirchenbauer and his colleagues have created an excellent work contributing to the body of knowledge, and the paper deservingly won an award for the oral performance the team behind it gave at the ICML 2023 conference. The paper is eloquent but direct, making it approachable for a broad audience, catering well to those outside the circles of Machine Learning.

The work's arguments are clear, evidence-backed, and examples are given to make clear how the model functions and how its outputs appear. In particular, the special care given to explanations of the two critical hyperparameters, named gamma and delta, are thorough and ensure understanding.

Even with many strong suits, the paper does carry with it a number of assumptions that lead to certain aspects of explanation being less clear. For example, its discussion of the impacts of beam search on algorithm performance leaves no discussion of what beam search is or does, which is a strangely high-level assumption given the approachable nature of the rest of the text. This is fortunately remedied with some quick side research, as beam search is not a complex topic.

4. Second Paper: On the Reliability of Watermarks for Large Language Models

4.1. Storyline

Motivation In the majority of cases, steganographic methods have been presented and evaluated only in ideal cases, where text remains unmodified and is available to be processed without any realistic complications to detection. The second work by Kirchenbauer et al. describes a more robust testing methodology that mimics common real-world scenarios with machine-generated text (Kirchenbauer et al., 2023b). These ”attacks” on identifiability include machine synthesis/replacement, copy-and-paste, and even human paraphrasing in a controlled setting, all of which are feasible approaches by which a dishonest student, an influencing campaign, or a fully-generated blog might go about evading detection.

Prior Work Prior work in the “under attack” setting is sparse, but the team compares watermarking approaches with more classical methods of machine generated text detection, like post-hoc classification with a trained model. In addition, the far more recent addition of so-called ‘Retrieval’ algorithms to the field are considered. Kirchenbauer also compares with a more adept version of their original watermark detector which is better equipped to detect bursts of watermarked content in larger-form settings.

Research Gap Previous works have failed to adequately strain methods against real-world settings for machine generated text, leading to optimistic overestimates of performance. Human paraphrasing, machine replacement algorithms, and copy-paste into long-form content are all common vectors of watermark and identity degradation that must be thoroughly addressed to prove the validity of watermarking capabilities.

Proposed Contributions This work proposes that Watermarking, and its newer, older brother WinMax Watermarking, perform sufficiently well under a variety of adversarial conditions that are used in the wild. These techniques also outperform competitors in the DetectGPT and Retrieval algorithms in the average case, in addtion to scaling far better over increasing token counts.

4.2. Proposed Solution

Kirchenbauer and his coauthors come to this conclusion by:

- Optimizing the watermark generation and detection systems for use in realistic scenarios.
- Studying the reliability of watermarking against paraphrasing by strong LLMs such as GPT 3.5.
- Testing “Copy-and-paste” style attacks where watermarked and otherwise machine-generated text is injected into larger human-generated text media.
- Conducting a human study where participants rewrite and paraphrase watermarked text with the explicit intention of removing the watermark.
- Comparing reliability estimates of Watermarking with other methods, such as loss-based and retrieval-based detection schemes, to contrast performance strengths.

This work serves as a substantially more data-driven companion paper to Kirchenbauer's first work on watermarking, discussed as the primary paper herein.

4.3. Claims/Evidence

Primary Claim The most substantial claim this work makes is that the proposed watermarking method, when compared with the DetectGPT classifier and the Retrieval method from (Krishna et al., 2023), performs markedly better in the robustness and reliability scenarios developed to mimic adversarial conditions for hiding machine-generated text.

Primary Evidence This claim is evident in Figure 8 of the work, which demonstrates higher accuracy for WinMax and Z-Score watermark testing in all but the GPT replacement attack where Z-score testing is matched by Retrieval. True positive rate remains high while false positive rates remain low in the subfigures, indicative of high performance both relatively *and* generally in these test environments.

Secondary Claim Also introduced in the work is the WinMax detection scheme, which the team claims is superior to the original Z-testing method. Z-score testing is computed globally, making it vulnerable to insertion into human-generated text and insensitive to segments of machine-generated text within a wider scope, thus limiting its best use case to text already suspected of foul play. WinMax instead uses a windowed statistic test over the tokens, able to effectively identify subsets of text that are highly probable to have originated from machine generation.

Secondary Evidence Figures 8 and 9 in the work display superior performance in nearly all cases for WinMax over Z-score testing with identical watermarking setups. As WinMax does not contain a substantial amount more overhead in testing that the Z-score approach, the comparison is fair and proves WinMax a more robust detection scheme.

Tertiary Claim The final claim from the work is that performance as a function of token count is a necessary metric for benchmarking steganographic methods for machine text generation, as they display unique dependencies on token quantity.

Teritiary Evidence Figure 9 displays that the two watermarking schemes continue to trend in a positive direction as token count increases, whereas Retrieval accuracy and *especially* DetectGPT accuracy can begin to decrease as token count continue to grow, particularly past ~ 700 tokens.

4.4. Critique/Discussion

Although by the same group, Kirchenbauer et al.'s work on the reliability of watermarking and other methods in realistic settings is a denser piece than the first. It is, however, still quite approachable and well written for a sufficiently broad audience.

The insights of the WinMax optimization were particularly intriguing, and the performance of DetectGPT displayed in the data collected is curiously poor for a state-of-the-art method. This on its own appears to validate the claims of the work that current approaches are applied in situations too ideal to create an adequate picture of relative quality.

5. Third Paper: Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation

5.1. Storyline

Motivation Hiding information within text, images, and video can be both useful and important. Researchers over 20 years ago knew this was a useful technique. However, discrete data like text is far harder to apply an encoding scheme to than continuous data such as color in an image or video. (Atallah et al., 2001) were working on these types of steganographic methods before the age of machine learning had begun, focusing instead on a more general concept of embedding ownership within something that is equally undetectable and difficult to remove.

The same principles still apply to the modern context of transformers and LLMs, as watermarking methods still attempt to differentiate 'ownership' between human and machine.

Prior Work Prior work on this time focused on the theory behind steganography and its simpler applications to continuous data, especially from the earliest instances of the Information Hiding Conference. The work also references a number of sources on pure natural language breakdown. The division of CERIAS at Purdue from which this paper was born is a linguistics and natural language processing lab, and the work focuses heavily on ontology, natural language analysis, and syntactic nuance.

The pre-softmax layer of modern LLMs proves hugely beneficial in blackboxing the nuance of the English language, and thus these techniques are discarded in favor of the simpler stochastic methods described earlier.

Research Gap The research gap observed was one of general text watermarking schemes, particularly on the implementation front, as unifying the cryptographic and linguistics research for a brute-force or elegant algorithmic solution required a vast breakdown of text content to maintain quality and utility.

Proposed Contributions The work proposes a novel scheme for an in-text watermark generation method, taking advantage of text-meaning-representation trees and a set of meaning-invariant transformations to change text content but maintain text quality.

5.2. Proposed Solution

The proposed scheme for encoding consists of three steps iterative steps to encode all watermark bits:

- Pick a valid, unmarked sentence of the lowest rank.
- Apply β transformations to encode β bits of data in the sentence.
- Re-sort the modified sentence in the set of sentences

5.3. Claims/Evidence

Primary Claim The primary claim of the work is that the proposed syntactic tree-based system works as a watermark for text input.

Primary Evidence Section 4.1 expands the iterative algorithm, providing a method by which to encode all bits of the desired watermark in a controlled and recoverable manner through Adjunct Movement, Clefting, and Passivization transformations.

Secondary Claim In addition to functioning as a watermark, the paper claims that the algorithm does not damage text utility.

Secondary Evidence Figure 2 provides evidence that the Text-Meaning Representation Trees used in the watermarking algorithm can be used as a framework to perform textual substitution with the Parser and replacement algorithms, all the while without degrading the sentence structure, thereby maintaining full text utility while encoding the watermark.

Tertiary Claim Lastly, the work claims that false positive rates for this exceedingly small for any reasonable watermark size.

Teritiary Evidence The work encourages padding watermark length with additional characters, recommending more than 12 extra in the provided example in section 4.4. As the false positive rate is determined to be $2^{-\omega}$, the math follows that $2^{-12} \approx 0.02\%$ as a base chance for any watermark, falling exponentially per added watermark character, is sufficient to make the chance of a false positive vanishingly small.

5.4. Critique/Discussion

This work is incredibly dense. Its expectations of natural language and linguistics area knowledge, as well as more general concepts such as trees, quadratic residues, and set notation, make comprehension challenging. Its lack of figures beyond tree diagrams place the burden of most visualization on the reader.

Behind the steep knowledge requirements is a fascinating and thorough work that explores some unique and cleverly positioned aspects of language structure to create a robust and complex encoding scheme that manages the difficulty of maintaining text utility. The techniques invented by the group are beyond impressive.

I still have questions concerning some of the inner workings of the algorithm and how replacements and transformations are carried out with the parser (which in itself seems an endlessly complex object). However, the use of prime numbers as the secret to the encoding and working to understand the problem from a watermarking perspective rather than a linguistics or programming one makes approaching the scheme easier.

6. Implementation

6.1. Motivation

As LLMs grow more pervasive and more powerful, having intense familiarity with their structure, inner workings, and deployment will become a more and more valuable skill. I find LLMs very interesting and would learn more through exploration of this project.

I'm also interested in how ML researchers are working to minimize the malicious capabilities of LLMs, which will become an ever-larger threat over the coming decades.

If my experiments are successful, I will hopefully be able to create a binary comparison of how effective watermarks are on multiple LLM systems. I hope to also demonstrate the use case of the private key functionality in addition to the re-implementation/environment deployment I've already completed.

6.2. Plan and Setup

The largest hurdle for this project is getting the provided codebase working. From there, the plan followed recreating the primary paper's experiments and findings, corresponding to a number of tests and hyperparameter modifications. These include:

- Evaluating the base watermark settings on provided input. (Terrapins)
- Evaluating the base watermark settings on custom input. (Moose, Transistors)
- Evaluating the base watermark settings on low-entropy input. (Milky Way)
- Performing hyperparameter tuning tests on custom input. (High delta, low gamma, high gamma)

- Evaluating the base watermark settings on a second model. (GPT-2)
- Performing attacks on watermarked text and evaluating watermark endurance. (Human Synthesis, GPT-4)
- Creating a custom watermark detector for watermarked subset detection. (Sliding Detection)

The custom watermark detector is code that will be written for this project, extending original codebase classes. Successful experiments and experimental metrics will be based upon the z-score and p-value metrics produced by the detector.

6.3. Details

Results gathered for this project were achieved primarily by using the codebase provided alongside Kirchenbauer et al.'s *A Watermark for Large Language Models*. Although the paper's original results were created using Facebook OPT-1.3, the results here are generated using OPT-6.7b and supplementary results for comparing watermarking across open-source LLMs use GPT-2. For mimicking attack vectors and typical use cases of LLM text, GPT-4 was used as a rephrasing engine, alongside typical human copy-paste and token substitution-style attacks.

6.4. Results

Parameter	Unwatermarked	Watermarked
Tokens	199	199
Greenlist Tokens	48	125
% Greenlist	24.1%	62.8%
Z-score	-0.286	12.3
p-value	0.613	$3.57 * 10^{-35}$
Prediction	Human	Machine
Confidence	H^0	100.000%

Table 1. Detection Statistics for OPT 6.7b Outputs

The non-watermarked and watermarked text gave two separate responses to the provided project prompt, which asked the LLM about the diamondback terrapin species. Without a watermark, it gave the following response:

- "… protected under the Endangered Species Act.

  The diamondback terrapin is the largest species in the family Terapinidae. It is a medium-sized terrestrial turtle, … may have a black tip. The diamondback terrapin has a pair of powerful jaws that are used for …

With a watermark, it instead produced the following:

- "… protected in the United States and Canada. The diamondback terrapin is the largest species in the family Terapinidae, and has been the most abundant terrestrial vertebrate in Chesapeake Bay for most of the past century. … loss of freshwater habitat can also have a direct impact on the availability of feeding resources. The loss of …

The full output can be found in Appendix A.1.

With tuned parameters and a high-entropy prompt (many input tokens), the detector was incredibly confident in identifying the watermarked sequence as watermarked, as well flagging the unwatermarked sequence in the Human/unwatermarked category. Not only does the detector work, but it runs in semi-reasonable time (500s) on an Nvidia 3070ti.

The same types of results can be seen in Appendix A.2 and A.3, corresponding to long prompts about Moose and Transistors. Long prompts were used at the recommendation of the paper, as OPT-6.7b was trained for such completions, but the Milky Way prompt in Appendix A.4 demonstrates that low-entropy prompts perform just as well for OPT.

Thereafter, hyperparameter tuning was verified on OPT. The first parameter, delta, corresponding to the strength of the soft watermark, was increased almost fourfold in the example in Appendix A.5.2. "Hardening" the watermark narrows the vocabulary breadth that the LLM is capable of using (in a distribution/probability fashion), corresponding to a more noticable watermark for both humans and the detector. A p-value of 10^{-106} demonstrates this quite clearly.

Second, exploring a low gamma, making for a small greenlist proportion, yielded a more constricted writing style as the LLM has fewer words to choose from to seed more often than average. With the Transistors prompt, this leads to it favoring certain words and topics while strongly avoiding others.

Lastly, a high gamma was introduced, making the greenlist large relative to the redlist. In order to generate a significant statistic, the greenlist had to be favored more often, as the final output had to have a high proportion of greenlist tokens to substantially deviate from the 50% unwatermarked baseline.

Both gamma adjustments passed both the statistic and eye tests for robustness and readability, but the researchers were spot on with their choices.

Following this, a subset of the OPT tests were performed on a GPT-2 model as well. They yielded similar results, though GPT-2 was more susceptible to odd language coming out of a modified delta parameter.

The attacks on the text yielded fantastically surprising results. The paper and its follow-up discussing reliability and robustness indicated decent performance against replacement and copy-paste attacks, but the results of those tests were stunning. All attacks were unsuccessful in erasing or degrading the watermark, including the full replacement attack vector through GPT-4.

Lastly, the new detection scheme mentioned in the implementation plan was developed and tested. It consists of a sliding-window detection, corresponding to detecting subsets of generated text within a larger document. With the provided examples, it gives full detection but also highlights subsections of the evaluated text for detection matches. Token groups in the index ranges *(150-170), (160-180), (220-240), (230-250), (260-280), (125-175), (150-200), (175-225), (200-250), (225-275), (75-225), (150-301),* and *(0-301)* all triggered the detector to flag watermarking, showing a bias towards the second half of the submitted text, which is the machine-generated segment.

7. Conclusions

Throughout this work, open-source LLMs OPT 6.7b and GPT-2 were used to evaluate the efficacy of the proposed soft watermarking methodology for machine generation detection. The basis of watermarking in machine text, past methods for watermarking and steganographic methods, and the newest proposed approach for watermarking text discreetly and robustly are evaluated. These models and the watermark engine are put to the test in realistic environments and prove an ability to watermark with extreme resistance to false positives and negatives, as well as high robustness to practical attack vectors.

For the purposes highlighted in the motivation of such a technique, notably cheating, influencing campaigns, and mass media generation, soft watermarks appear a plausible and useful scheme to rein in some of the rampant issues with language models today. It performs substantially better in text utility and perplexity than previous ”hard” watermark methods, corresponding to more helpful and unimpacted outputs. Soft Watermarking also addresses many of the ethics concerns with post-hoc methods like DetectGPT, namely disproportionate false positive rates against certain groups of human writers, such as non-native English speakers.

In its performance, this watermarking engine also proved to be a slim, low-overhead addition to generation procedures. Detection is incredibly quick due to its nature as a simple z-statistic test, and watermarking only modifies a single distribution based on its internal greenlist/redlist rules. There are no indications that implementation on a large scale would be of any substantial difficulty or degrade performance. The robustness of the watermark is highly impressive, resisting even an entire rephrasing by the a state-of-the-art gated-access model. If an average student was told that their work was watermarked, this is what they would likely try to correct it at a low-effort cost, and it is show here to *not work*.

This is not to say that this watermark, or any watermarking schema for that matter, solves the issues presented by the decade of the LLM. The main drawback of watermarking is that it needs to be embedded at generation time, a problem not shared by DetectGPT and other post-hoc detectors. Teams responsible for ChatGPT, Bard, Llama, and other systems do not have strong incentives to burden their systems with additional layers, though the extreme scrutiny with which these groups are viewed may incline them to add watermarking to their models.

Required alongside a watermarking engine is a watermark detector, which presents its own set of challenges, most notably in the realm of security. As the engine-detector pair rely on secret parameters and hashing for the model to work, sufficient volumes of detection checks can clue an attacker in on the possible hash and degrade the security of the watermark. Thus the detector must be a limited-access, behind-API elements, and there are additional tradeoffs to consider in the level of detail provided to the end user (say a college-level english professor checking students' essays). Providing detected token counts, z-scores, p-values may inspire more confidence in the tool, but also would make cracking the hash easier.

The final challenge is the human element. LLMs, even open source versions, are already prevalent enough without watermarks that intelligent attackers may be unaffected. If students began to hear that GPT-4 had cheat detection capabilities enabled, many are proactive enough to switch to Bard, to Llama, to download one, or even train their own from a base pretrained model to bypass these techniques. OPT 6.7b is too computationally demanding for many users and many applications on a home implementation, and GPT-2 produces text of too poor a quality to pass as perfectly human, but there will undoubtedly be more models released that add to the growing cat-and-mouse game of machine-generated text detection.

References

Atallah, M. J., Raskin, V., Crogan, M., Hempelmann, C., Kerschbaum, F., Mohamed, D., and Naik, S. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In Moskowitz, I. S. (ed.), *Information Hiding*, pp. 185–200, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-45496-0.

Johnson, M., Davidson, S., Eschbach, D., and Watson, C. Gene sequence inspired vhdl plagiarism screening. *American Society for Engineering Education*, Jun 2004. doi:

10.18260/1-2--13875.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., and Goldstein, T. On the reliability of watermarks for large language models, 2023b.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, 2023.

A. OPT-6.7b Unabridged Experimental Outputs

The complete outputs (which were terminated at 199 tokens for OPT 6.7b) are shown here. Both watermarked and unwatermarked outputs display adequate readability, and it can be conceived how a more advanced model like GPT-4 could bury the watermark even better. The full prompts for each test are provided as well.

A.1. Diamondback Terrapins

A.1.1. Prompt

"The diamondback terrapin or simply terrapin (Malaclemys terrapin) is a species of turtle native to the brackish coastal tidal marshes of the Northeastern and southern United States, and in Bermuda.[6] It belongs to the monotypic genus Malaclemys. It has one of the largest ranges of all turtles in North America, stretching as far south as the Florida Keys and as far north as Cape Cod.[7] The name 'terrapin' is derived from the Algonquian word torope.[8] It applies to Malaclemys terrapin in both British English and American English. The name originally was used by early European settlers in North America to describe these brackish-water turtles that inhabited neither freshwater habitats nor the sea. It retains this primary meaning in American English.[8] In British English, however, other semi-aquatic turtle species, such as the red-eared slider, might also be called terrapins. The common name refers to the diamond pattern on top of its shell (carapace), but the overall pattern and coloration vary greatly. The shell is usually wider at the back than in the front, and from above it appears wedge-shaped. The shell coloring can vary from brown to grey, and its body color can be grey, brown, yellow, or white. All have a unique pattern of wiggly, black markings or spots on their body and head. The diamondback terrapin has large webbed feet.[9] The species is"

A.1.2. Unwatermarked

"protected under the Endangered Species Act.

The diamondback terrapin is the largest species in the family Terapinidae. It is a medium-sized terrestrial turtle, usually reaching a length of about 55 cm (22 in) and weighing up to 220 g (8 oz). Females are generally larger than males. Its carapace is generally dark brown or grey. The head and body are usually yellowish or white. It has small webbed feet with rounded or pointed tips. The diamondback terrapin has a distinctive diamond pattern on its head and neck. The upper lip of its mouth is black. It has a yellow or golden iris. The diamondback terrapin has a long, slender neck. It usually has a single red eye with a white ring around it. Its tail is long and slender, and is usually yellowish or white. It may have a black tip. The diamondback terrapin has a pair of powerful jaws that are used for"

A.1.3. Watermarked

"protected in the United States and Canada.

The diamondback terrapin is the largest species in the family Terapinidae, and has been the most abundant terrestrial vertebrate in Chesapeake Bay for most of the past century. The decl ine in the abundance of the species was not known until 2002, when the turtles began to decline. The decline was attributed to a combination of factors. The loss of riparian habitat was noted in the mid-1980s, and this habitat was lost when the Bay Bridge was constructed in the mid-1990s. The loss of riparian habitat can also have a direct impact on the availability o f feeding resources. The loss of freshwater habitat was noted in the mid-1980s, and this habitat can also be lost when a dam is removed, as was done when the Caesar Rodney Dam in Dover, Delaware was removed in the mid-1990s. The loss of freshwater habitat can also have a direct impact on the availability of feeding resources. The loss of"

A.2. Moose

A.2.1. Prompt

"The moose or elk is the only species in the genus Alces. The moose is the tallest and second-largest land mammal in North America, only falling short of the American buffalo in terms of mass. It is the largest and heaviest extant species of deer. Most adult male moose have distinctive broad, palmate antlers; most other members of the deer family have antlers with a dendritic configuration. Moose typically inhabit boreal forests and temperate broadleaf and mixed forests of the Northern Hemisphere in temperate to subarctic climates. Hunting and other human activities have caused a reduction in the size of the moose's range over time. It has been reintroduced to some of its former habitats. Currently, most moose"

A.2.2. UNWATERMARKED

“are found in North America and Canada.

The moose is the largest living member of the deer family. Adult males have antlers, while adult females have no antlers. The antlers are typically broad and pointed, although in some cases the horns are narrow and pointed. The antlers of a mature male usually grow from October to February. An adult female typically grows antlers from December to February.

Moose are the largest land-dwelling mammal in North America. The largest recorded moose was a female moose that weighed approximately 3,000 lb (1,361 kg). The moose has the largest antlers in the animal kingdom; a bull’s antlers can measure up to 6.5 ft (2 m) in length, and weigh up to 230 lb (105 kg) in mass. The antlers can grow to a maximum height of 9 ft (2.7 m) and a maximum weight of 1,800 lb (814 kg”

This text was found to have 60 of its 199 tokens within the greenlist, corresponding to a Z-score of 1.68 and a p-value of 0.0467.

A.2.3. WATERMARKED

“ live in the boreal and tundra areas of North America. The moose was domesticated in the 14th century.

The male’s antlers (antelopes) grow each spring and grow out again after the antler sheds in the fall. The antlers are made of bone, cartilage, and skin. The male’s antlers start off as a tuft of hair and end in the form of a sharp point that can be used for fighting or for attracting a mate.[2] Some males have horns that are used for fighting as well. The antlers of the male’s antlers are usually as much as 2–2.5 m (7.3–8.4 ft) long. The horns of the male’s antlers start off as a tuft of hair and end in the form of a sharp point that can be used for fighting or for attracting a mate.[2]

The female moose (alces alces) is ”

This text was found to have 105 of its 199 tokens within the greenlist, corresponding to a Z-score of 9.04 and a p-value of $7.49 * 10^{-20}$.

A.3. Transistors

A.3.1. PROMPT

“A transistor is a semiconductor device used to amplify or switch electrical signals and power. It is one of the basic building blocks of modern electronics.[1] It is composed of semiconductor material, usually with at least three terminals for connection to an electronic circuit. A voltage or current applied to one pair of the transistor’s terminals controls the current through another pair of terminals. Because the controlled (output) power can be higher than the controlling (input) power, a transistor can amplify a signal. Some transistors”

A.3.2. UNWATERMARKED

“are designed to switch electrical power on and off.

The first transistor was a vacuum tube, a device used to amplify radio signals in the 1920s.[1] The first transistor was designed by the German Friedrich Nennig in 1936.[1]

There are two main types of transistor: field effect transistor (FET) and bipolar transistor. A FET is a device that controls current flow by controlling the voltage between two terminals (gate and source) of the FET. A bipolar transistor is a device that amplifies the small voltage across its terminals by creating a current loop that flows from the base to the emitter and then back to the base.

The transistor was invented by German engineer Friedrich Nennig and was first described in a paper in the journal Zeitschrift für Elektrotechnik.[1]

Contents

The transistor was invented by German engineer Friedrich Nennig, who was working at Siemens in Berlin at the beginning of”

This text was found to have 56 of its 199 tokens within the greenlist, corresponding to a Z-score of 1.02 and a p-value of

0.153.

A.3.3. WATERMARKED

"are designed to switch electrical power on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the control voltage, and in the case of a bipolar transistor, on the current base and collector current). A control electrode (gate) controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is "off" and there is "no current" in the device). As the control electrode is increased in voltage, the current increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either "off" (connected) or "on" (not connected). A half-bridge (two transistors) or "

This text was found to have 106 of its 199 tokens within the greenlist, corresponding to a Z-score of 9.21 and a p-value of $1.65 * 10^{-20}$.

A.4. Low Entropy Prompt - Milky Way

This prompt, being much shorter, was an experiment in testing OPT 6.7b's capabilities more than anything, as the watermarking was completely unaffected with a much shorter prompt length. Entropy issues arise for watermarking when *detection* is run on short-length text, but I was curious what would happen. Most notably, the model took around 50s to produce an output rather than 500. Neither run seemed to understand how the Milky Way got its name!

A.4.1. PROMPT

"The Milky Way is the galaxy that includes the Solar System, with the name describing the galaxy's appearance from Earth: a"

A.4.2. UNWATERMARKED

"spiral. It's the third biggest galaxy in the Universe. The Local Group consists of the Milky Way, Andromeda Galaxy, Magellanic Clouds, and three smaller galaxies. Thanks. I've always thought of the Milky Way as the galaxy the solar system is in, not the whole galaxy."

This text was found to have 21 of its 59 tokens within the greenlist, corresponding to a Z-score of 1.88 and a p-value of 0.0301.

A.4.3. WATERMARKED

"spiral. The galaxy is the largest and most massive of the Milky Way's three superclusters, and contains about 200 billion stars. The galaxy is orbited by approximately thirty thousand solar systems and has an average distance from Earth of approximately 120,000 light-years. The galaxy is thought to have a mass of about 400 billion solar masses. The galaxy was named after the Ancient Roman god of the Milky Way, and the Milky Way Galaxy is the largest visible object in the night sky.

Contents

The Milky Way Galaxy was the first celestial object to be observed by a human. The first known to have seen it was the Ancient Greek astronomer Hipparchus. The galaxy was observed by Tycho Brahe in 1604 and by William Herschel in 1784.[2]

It was discovered by Matthias Johann Bayer in 1609. The first image of the Milky Way, comparing Bayer's discovery to the Milky Way as seen from Earth, was taken by Johann Bayer and his"

This text was found to have 117 of its 199 tokens within the greenlist, corresponding to a Z-score of 11.0 and a p-value of $1.72 * 10^{-25}$.

A.5. Hyperparameter tuning

In these runs, hyperparameters gamma and delta are adjusted to observe impacts on text quality, detectability, and verify the primary paper's claims.

A.5.1. PROMPT

The same prompt from the *Transistors* section is used.

A.5.2. $\delta = 7$

"are designed primarily as switching elements. They are used in switching power amplifiers, for instance radio transmitters.[2] Although the basic building blocks for transistors were first made by J.J. Thomson towards the end of the 19th century,[6][9] the electronic transceiver using the "thyristor", a device built by RCA in 1935, was not commercially viable due in no small measure because of the prohibitively high costs.[7][10] To create this first transistor, all components were built from scratch in the laboratory. The Cornell circuit was used for its simplicity of design, and for its simplicity of building. The Cornell circuit was used in the RCA 3060 transmitter in 1935. The transistor, however "routine" in the laboratory is the key word here. The transistorm was not commercial. The transistor, however "routine" in the laboratory is the key word here. The transistor, however "routine" in the laboratory"

This text was found to have 183 of its 199 tokens within the greenlist, corresponding to a Z-score of 21.8 and a p-value of $1.72 * 10^{-106}$.

This is an almost unfathomably small probability, but the text quality is entirely compromised as a result of the harder watermark pulling almost exclusively from the greenlist. This clearly causes some suboptimal generation and is therefore not a beneficial value for the hyperparameter.

A.5.3. $\gamma = 0.1$

"Theory

Theory of operation

A transistor is a device that can be used in the electronic, radio-frequency and optical fields. It consists of a PN junction (a diode) between two semi-conductors of opposite conductivity types, and has an internal current gain that can be in the range of 1,000 to 10,000. The current gain depends on the proper application of the transistor, and refers to the ratio between the current gain the collector current to the emitter current. The current gain can be changed by changing the emitter and/or base resistances. The transistor is characterised by its heterojunction, its base-emitter voltage, and its base current gain.

Types

Voltage regulators

Voltage regulators, often called power amplifiers (PA), are electric devices that regulate an electric current to a level within a specified range, for"

The text is still very readable, even with such a tiny greenlist. It does read as slightly constricted writing, but may be a symptom of confirmation bias.

A.5.4. $\gamma = 0.5$

"are designed for switching and are used in switches and power supplies. The field-effect transistor (FET) family of transistors includes some of the most common transistors in modern electronic circuit design and power electronics.[2] A transistor is the fundamental building block of the electronic circuit, and therefore of the electronic device itself. Its operation can be described by comparing the control electrode's current with the control electrode's voltage—the difference between the voltages creates the control current, or voltage-controlled current, that changes the current through the base.[3] Theory The transistor can work in three different operating regions. It is commonly called the base-emitter (BE) region, base-collector

(BC) region and emitter-collector (EC) region. The BC region is the region in which electrons (or holes) move into the base material. When a voltage is applied to the control electrode (gate), the electrons (or holes) move to the"

The text is still entirely readable and passes the z=4 test even with a larger base greenlist size.

B. GPT-2 Unabridged Experimental Outputs

The complete outputs (which were also terminated at 199 tokens for GPT-2) are shown here. GPT-2 is an earlier transformer model than OPT 6.7, so its textual outputs are substantially less human-passing. One benefit to GPT-2 is that its smaller model size makes generation immensely faster than OPT 6.7 - down to around 17 seconds from 500.

B.1. Transistors

B.1.1. Prompt

"A transistor is a semiconductor device used to amplify or switch electrical signals and power. It is one of the basic building blocks of modern electronics.[1] It is composed of semiconductor material, usually with at least three terminals for connection to an electronic circuit. A voltage or current applied to one pair of the transistor's terminals controls the current through another pair of terminals. Because the controlled (output) power can be higher than the controlling (input) power, a transistor can amplify a signal. Some transistors"

B.1.2. Unwatermarked

" have an equal or greater level of control, while others have a lower level of control.

The 3D transistor is a direct connection between a transistor and an electronic circuit. Because the transistor acts as a direct link between two parts, it is often called a sub-section of an electronic circuit. The 3D transistor also has two independent sub-sections, called the "3D junction" or the "V-shaped" junction. The 3D transistor is the only semiconductor that can be connected to an electronic circuit. The 3D junction is the center of the 3D transistor.

3D transistor is the simplest in its class, used to amplify or switch electrical signals and power. It is a semiconductor device used to amplify or switch electrical signals and power. It is one of the basic building blocks of modern electronics.[1] It is composed of semiconductor material, usually with at least three terminals for connection to an electronic circuit. A voltage or current applied to one"

This text was found to have 52 of its 199 tokens within the greenlist, corresponding to a Z-score of 0.368 and a p-value of 0.356.

B.1.3. Watermarked

"have an output voltage of 3.5 volts, and can vary between 2.5 volts and 6.4 volts depending on the operation situation. For example, the current in a 6.4-volt switch is near 6.4 volts, and can vary from 4.2 to 4.6 volts depending on the power output. The voltage of the switch determines the current in the circuit and the current in the circuit's output. The current in a 6.4-volt USB circuit can vary by as much as 45 percent in a 20-watt (40-watt) light bulb.

As with many other devices in electronics, the 'dumb' transistor can also serve as an output amplifier or a motor or an inductor. The diode, which is used as an output for the output circuit, can also be an inductor, which acts as an interface device. This is because the inductor is an inductive current current circuit that links the two currents."

This text was found to have 135 of its 199 tokens within the greenlist, corresponding to a Z-score of 14.0 and a p-value of $1.44 * 10^{-44}$.

These text blocks fail to match the scientific writing style of the prompt, but still yield very readable outputs. There is no significant difference noticable between the two in terms of quality or readability, indicating proper extendability of the watermarking scheme to other models.

B.2. Transistors with Hyperparamter Tuning

B.2.1. PROMPT

The same prompt from the *Transistors* section is used.

B.2.2. $\delta = 7$

" that have this power only allow one terminal, and it is not known which of the two terminals are in the same current, that will be needed in a transistor's output. The current in a transistor's output usually occurs at two different points in the circuit: at subvoltage (DVT), during low phase, and also when the current in the circuit can not exceed the current in the circuit's subvoltage (DVT). In certain embodiments an output circuit can include power from multiple terminals. For example: when the current in the circuit's subvoltage (DVT), is at 2 or less (and is less than 1 Hz constant, depending how high any of the current flows throughout an output circuit), this power will be needed in the circuit's subvoltage (DVT). In certain embodiments an output circuit can include power from multiple terminals. For example: when the current in the circuit's subvoltage (DVT), is at 2 or less (and is less"

This text was found to have 193 of its 199 tokens within the greenlist, corresponding to a Z-score of 23.5 and a p-value of $6.41 * 10^{-122}$.

This output represents proper English, but not much else. Such a high watermark strength severely degrades the output quality on this older model, and causes it to go in circles. The overpresence of phrases "*In certain embodiments*" and "*subvoltage (DVT)*" give away something being seriously wrong with the text.

C. OPT-6.7b Watermarked at base settings vs. various attack vectors

These experiments use the watermarked output from Appendix A.3.3. and modify it along different attack strategies.

C.1. Original Watermarked Text

"are designed to switch electrical power on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the control voltage, and in the case of a bipolar transistor, on the current base and collector current). A control electrode (gate) controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is "off" and there is "no current" in the device). As the control electrode is increased in voltage, the current increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either "off" (connected) or "on" (not connected). A half-bridge (two transistors) or "

The z-score here is **9.21**, which passes the detection threshold.

C.2. Pasted/Concatenated Text

Here, we concatenate the prompt with the output, as a malicious actor might do when trying to complete a prompt.

"A transistor is a semiconductor device used to amplify or switch electrical signals and power. It is one of the basic building blocks of modern electronics.[1] It is composed of semiconductor material, usually with at least three terminals for connection to an electronic circuit. A voltage or current applied to one pair of the transistor's terminals controls the current through another pair of terminals. Because the controlled (output) power can be higher than the controlling (input) power, a transistor can amplify a signal. Some transistors are designed to switch electrical power on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the control voltage, and in the case of a bipolar transistor, on the current base and collector current). A control electrode (gate) controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is ”off” and there is ”no current” in the device). As the control electrode is increased in voltage, the current increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either ”off” (connected) or ”on” (not connected). A half-bridge (two transistors) or”

The z-score here is **7.42**. This passes the base detection threshold of 4.0, and as such the passage is correctly labelled as watermarked.

C.3. Replacement Attack

Here, certain text is modified and ”cleaned up” as a malicious actor might do to hide some of the idiosyncracies of machine generated text. Strikethroughs indicate deletions and italics indicate additions and modifications. There is no effort made to correct factual or logical errors in the output to mimic an uninformed attacker (i.e. a non-area expert).

“are designed to switch ~~electrical power~~ on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the *gate* voltage, and in the case of a bipolar transistor, on the ~~current~~ base and collector current. A control electrode*, the gate,* controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is ”off” and there is ”no current” in the device). As the control electrode *voltage is* increased~~ in voltage~~, the current *in the device* increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either ”off” (connected) or ”on” (not connected). A half-bridge (two transistors) or ”

The z-score here is **7.24**. This still passes the detection threshold with direct modifications to the generated text. This indicates the base parameters create a suitably robust watermark that is difficult to counteract.

C.4. GPT-4 Paraphrasing Attack

This final attack feeds the output into GPT-4, and the subsequent result into the detector. It is also given the last few words of the prompt to create a proper sentence at the beginning of the input.

“Some transistors are engineered to toggle the flow of electrical power.

Basic Principle

Operation Principle

A transistor comprises three terminals: two for linking to an outside circuit and one for regulating the current passing through the component.[2] The current is contingent upon the control voltage and, in the case of a bipolar transistor, on the base and collector current. A control electrode (gate) manages this current. When the control electrode maintains its lowest voltage, a flow moves from the emitter to the collector (the emitter is ”inactive,” and the device carries ”no current”). As the control electrode’s voltage rises, the current amplifies.[2]

Typically, a transistor functions as a single-pole, single-throw apparatus, signifying that its base is either connected or disconnected from an external circuit. The transistor operates in an ”off” state (connected) or an ”on” state (disconnected). A half-bridge configuration comprises two transistors, or”

The z-score here is **4.41**. This is a complete success of the watermark. GPT-4 replaced almost all of the functional text from an eye test, and it did not manage to erase the watermark.

C.5. GPT-4 Generation (For Comparison)

When asking GPT-4 to generate text on the subject of ”basic operating principles of transistors, it provided the following:

“Transistors are fundamental semiconductor devices crucial in modern electronics for amplification, switching, and signal modulation. They operate based on three main principles: amplification, modulation, and switching.

Amplification: Transistors amplify electrical signals. They consist of three layers—Emitter, Base, and Collector (in bipolar junction transistors, or BJT)—and operate in two main types: NPN and PNP. By applying a small signal at the base (in the case of a BJT), a larger current flows between the collector and emitter. This controlled amplification forms the basis of analog circuits.

Modulation: In modulation, transistors adjust an output signal based on variations in an input signal. For instance, in amplitude modulation (AM) radio, transistors modulate the amplitude of a carrier wave with the audio signal, enabling the transmission of sound over long distances.

Switching: Transistors function as switches in digital circuits. They operate either in an ”on” state (allowing current flow) or an ”off” state (blocking current). When used in digital systems, transistors control the flow of electricity, enabling the binary operations fundamental to computing.

Transistors revolutionized electronics, enabling miniaturization and paving the way for modern computing devices. Their ability to amplify, modulate, and switch electrical signals forms the basis of most electronic systems, from computers to smartphones, making them an indispensable component in today’s technology landscape.”

This text, which is not watermarked, gave a z-score of 1.16, within expected bounds for random text relative to the watermark.

D. OPT-6.7b with Modified Sliding Window Detector vs. various attack vectors

These experiments use the watermarked output from Appendix A.3.3. and modify it along different attack strategies in the same manner as Appendix C. The detector implemented is one that I modified to perform subset evaluations.

D.1. Original Watermarked Text

“are designed to switch electrical power on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the control voltage, and in the case of a bipolar transistor, on the current base and collector current). A control electrode (gate) controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is ”off” and there is ”no current” in the device). As the control electrode is increased in voltage, the current increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either ”off” (connected) or ”on” (not connected). A half-bridge (two transistors) or ”

The z-score here is **9.04**, which passes the detection threshold.

D.2. Pasted/Concatenated Text

“A transistor is a semiconductor device used to amplify or switch electrical signals and power. It is one of the basic building blocks of modern electronics.[1] It is composed of semiconductor material, usually with at least three terminals for connection to an electronic circuit. A voltage or current applied to one pair of the transistor’s terminals controls the current through another pair of terminals. Because the controlled (output) power can be higher than the controlling (input) power, a transistor can amplify a signal. Some transistors are designed to switch electrical power on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the control voltage, and in the case of a bipolar transistor, on the current base and collector current). A control electrode (gate) controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is ”off” and there is ”no current” in the device). As the control electrode is increased in voltage, the current increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either ”off” (connected) or ”on” (not connected). A half-bridge (two transistors) or”

The z-score here is **7.42**. This passes the base detection threshold of 4.0, and as such the passage is correctly labelled as watermarked.

D.3. Replacement Attack

“are designed to switch ~~electrical power~~ on and off.

Theory

Theory of operation

A transistor consists of three terminals: two for connection to an external circuit and one for controlling the current through the device.[2] The current depends on the *gate* voltage, and in the case of a bipolar transistor, on the ~~current~~ base and collector current. A control electrode*, the gate,* controls the current. When the control electrode is at its lowest possible voltage, a current flows from the emitter to the collector (the emitter is ”off” and there is ”no current” in the device). As the control electrode *voltage is* increased~~ in voltage~~, the current *in the device* increases.[2]

A transistor is usually a single-pole, single-throw device, meaning its base is either connected to an external circuit or not connected to an external circuit. The transistor is either ”off” (connected) or ”on” (not connected). A half-bridge (two transistors) or ”

The z-score here is **7**. This still passes the detection threshold with direct modifications to the generated text. This indicates the base parameters create a suitably robust watermark that is difficult to counteract.

D.4. GPT-4 Paraphrasing Attack

“Some transistors are engineered to toggle the flow of electrical power.

Basic Principle

Operation Principle

A transistor comprises three terminals: two for linking to an outside circuit and one for regulating the current passing through the component.[2] The current is contingent upon the control voltage and, in the case of a bipolar transistor, on the base and collector current. A control electrode (gate) manages this current. When the control electrode maintains its lowest voltage, a flow moves from the emitter to the collector (the emitter is ”inactive,” and the device carries ”no current”). As the control electrode’s voltage rises, the current amplifies.[2]

Typically, a transistor functions as a single-pole, single-throw apparatus, signifying that its base is either connected or disconnected from an external circuit. The transistor operates in an ”off” state (connected) or an ”on” state (disconnected). A half-bridge configuration comprises two transistors, or”

The z-score here is **4.41**.

D.5. GPT-4 Generation (For Comparison)

When asking GPT-4 to generate text on the subject of ”basic operating principles of transistors, it provided the following:

“Transistors are fundamental semiconductor devices crucial in modern electronics for amplification, switching, and signal modulation. They operate based on three main principles: amplification, modulation, and switching.

Amplification: Transistors amplify electrical signals. They consist of three layers—Emitter, Base, and Collector (in bipolar junction transistors, or BJT)—and operate in two main types: NPN and PNP. By applying a small signal at the base (in the case of a BJT), a larger current flows between the collector and emitter. This controlled amplification forms the basis of analog circuits.

Modulation: In modulation, transistors adjust an output signal based on variations in an input signal. For instance, in amplitude modulation (AM) radio, transistors modulate the amplitude of a carrier wave with the audio signal, enabling the transmission of sound over long distances.

Switching: Transistors function as switches in digital circuits. They operate either in an ”on” state (allowing current flow) or an ”off” state (blocking current). When used in digital systems, transistors control the flow of electricity, enabling the binary operations fundamental to computing.

Transistors revolutionized electronics, enabling miniaturization and paving the way for modern computing devices. Their ability to amplify, modulate, and switch electrical signals forms the basis of most electronic systems, from computers to smartphones, making them an indispensable component in today’s technology landscape.”

This text, which is not watermarked, gave a z-score of 1.16, within expected bounds for random text relative to the watermark.