WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models

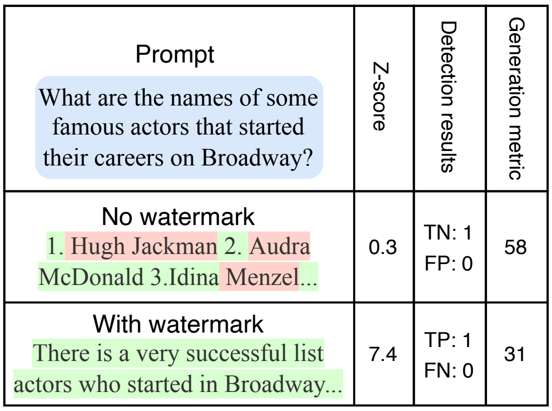
Shangqing Tu[1*], **Yuliang Sun**[2*], **Yushi Bai**[1], **Jifan Yu**[1], **Lei Hou**[1†] and **Juanzi Li**[1]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]School of Computer Science and Engineering, Beihang University
`{tsq22,bys22,yujf21}@mails.tsinghua.edu.cn`
`21371245@buaa.edu.cn,{houlei,lijuanzi}@tsinghua.edu.cn`

Abstract

To mitigate the potential misuse of large language models (LLMs), recent research has developed watermarking algorithms, which restrict the generation process to leave an invisible trace for watermark detection. Due to the two-stage nature of the task, most studies evaluate the generation and detection separately, thereby presenting a challenge in unbiased, thorough, and applicable evaluations. In this paper, we introduce WaterBench, the first comprehensive benchmark for LLM watermarks, in which we design three crucial factors: (1) For **benchmarking procedure**, to ensure an apples-to-apples comparison, we first adjust each watermarking method's hyper-parameter to reach the same watermarking strength, then jointly evaluate their generation and detection performance. (2) For **task selection**, we diversify the input and output length to form a five-category taxonomy, covering 9 tasks. (3) For **evaluation metric**, we adopt the GPT4-Judge for automatically evaluating the decline of instruction-following abilities after watermarking. We evaluate 4 open-source watermarks on 2 LLMs under 2 watermarking strengths and observe the common struggles for current methods on maintaining the generation quality. The code and data are available at `https://github.com/THU-KEG/WaterBench`.

1 Introduction

LLM has achieved significant success in generating human-like texts (Cai et al., 2023; OpenAI, 2023; Bubeck et al., 2023). However, the potential misuse of LLM has also raised concerns (Li et al., 2023a). For example, ChatGPT can be used to generate fake news (Wang et al., 2023b), which may manipulate the public opinion. To mitigate this kind of risk, it is necessary to develop a watermarking algorithm to detect whether a text is generated by LLM (Kirchenbauer et al., 2023a). As shown in Figure 1, the watermarked texts are generated with a biased distribution of tokens, which distinguishes it from unwatermarked texts. We believe the goal of watermarking is to achieve high detection accuracy while maintaining the generation quality. So we utilize the commonly used TP (True Positive), TN (True Negative), and GM (Generation Metric) to evaluate watermarks (Ghosal et al., 2023).



Prompt What are the names of some famous actors that started their careers on Broadway?	Z-score	Detection results	Generation metric
No watermark 1. Hugh Jackman 2. Audra McDonald 3.Idina Menzel...	0.3	TN: 1 FP: 0	58
With watermark There is a very successful list actors who started in Broadway...	7.4	TP: 1 FN: 0	31

Figure 1: The generated texts without and with watermark (Kirchenbauer et al., 2023a) on a test example from AlpacaFarm (Dubois et al., 2023), an instruction-following benchmark. LLM equipped with watermark will be more inclined to generate tokens in the green list, which can then be detected by a higher z-score measurement ($z > 4$). We utilize TP, TN, and GM to jointly evaluate the watermarking performance.

Due to the two-stage nature of this task, most studies (Kuditipudi et al., 2023; Zhao et al., 2023) evaluate the generation and detection separately and they do not conduct a unified hyper-parameter search for each watermarking method, which may lead to unfair comparisons. Since, there is usually a trade-off between the detection performance and the generation quality. Besides, previous evaluations are often conducted via text completion on a single dataset, such as C4 RealNewsLike dataset (Raffel et al., 2020), which cannot comprehensively measure the generation quality of LLMs.

---

[*] Equal Contribution.

[†] Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1517–1542
August 11-16, 2024

Research for Watermark	Control Hyper Para.	Jointly Test (TP,TN,GM)	Number of Tasks	Instruction Following	Metric for Generated Text
LM Watermark (Kirchenbauer et al., 2023a)	✓	×	1	×	Perplexity
V2 Watermark (Kirchenbauer et al., 2023b)	✓	×	2	×	Perplexity
Robust Watermark (Kuditipudi et al., 2023)	✓	×	1	×	Perplexity
GPT Watermark (Zhao et al., 2023)	×	×	2	×	Perplexity
Semantic Watermark (Fu et al., 2023)	✓	×	2	×	Ref.
Three Bricks (Fernandez et al., 2023)	✓	×	3	×	Ref.
WaterBench (ours)	✓	✓	9	✓	Ref./GPT4-Judge

Table 1: Comparison with existing works' evaluations of LLM watermarks. The column *Jointly Test* means whether the three metrics for each watermark are jointly tested under one run. The column *Instruction Following* means whether it evaluates this ability. The term *Para.* and *Ref.* are short for parameter and reference-based metric.

Furthermore, most evaluations only calculate the perplexity (Kirchenbauer et al., 2023b), which is not aligned with human preference and thus not practical in the era of LLMs (Chia et al., 2023).

To address these issues, we propose WaterBench, the first comprehensive benchmark for LLM watermarks, which has three crucial factors: (1) **Benchmarking Procedure**: We first introduce the concept of watermarking strength (Mei et al., 2002), i.e. the detection robustness to disturbance, to quantify the LLM watermarks' trade-off controlled by hyper-parameters. We present a reasonable hyper-parameter search procedure: Given a dataset and an LLM, we adjust the hyper-parameters of each watermarking method to unify the watermarking strength and then freeze the parameters to jointly evaluate the detection and generation performance. (2) **Task Selection**: To add disturbance on watermarks, we differentiate the task settings based on the length of input and output, which decides how much information the watermark can embed. Therefore, we form a new taxonomy with five task categories and nine sub-tasks, which are selected from existing datasets with various length settings (Dubois et al., 2023). (3) **Evaluation Metric**: We adopt the GPT4-Judge (Zheng et al., 2023) for automatically evaluating the instruction-following performance decline after watermarking. Then we conduct a human evaluation to verify the agreement between the human and GPT4.

Based on the WaterBench dataset, we conduct an experiment of 4 reproducible watermarks on 2 LLMs (Llama2-chat (Touvron et al., 2023) and InternLM (Team, 2023)), leading to some interesting findings: (1) We adjust two different watermarking strengths, 0.7 and 0.95, and observe that the detection and generation performance are significantly different. In other words, if we compare two watermark strategies without aligning their watermarking strengths, it is easy to let one "surpass" another in some aspects. (2) The tasks with short output length are generally more difficult to detect, with lower TP. The V2 watermark (Kirchenbauer et al., 2023b) is the best watermarking method in terms of GM. (3) On the open-ended task, if we use GPT4-judge to evaluate, the watermarked LLM will decrease over 96% from original LLM, which shows the sensitivity of the metric and indicates the common struggles of watermarks on maintaining the generation quality. In human evaluation, the GPT4 obtains over 0.6 Cohen's kappa coefficient with 3 annotators, achieving substantial agreement.

To summarize, our contributions are three-fold: (1) We propose a new benchmarking procedure that first search hyper-parameters for watermarks then jointly evaluate detection and generation performance to eliminate the unfair comparison between different watermarking strengths. (2) We construct a multi-task benchmark to facilitate future research. (3) We incorporate GPT4-Judge to evaluate the watermarked LLMs, which effectively reflects the decline of generation quality.

2 Related Work

To detect LLM-generated text, previous works (Tu et al., 2023; Guo et al., 2023; Mitchell et al., 2023) mainly explored the classifiers that distinguishes human and LLM-generated texts based on features. However, as LLMs are becoming more and more alike human, some classifiers may mistakenly recognize human as LLMs (Sadasivan et al., 2023).

In addition to black-box classifiers, recent approaches have also introduced white-box detection methods that inject watermarks into LLM generated texts (Tang et al., 2023; Yang et al., 2023; Liu et al., 2024). The inference-time watermarks (Pan et al., 2024) randomly split the vocabulary and ad-

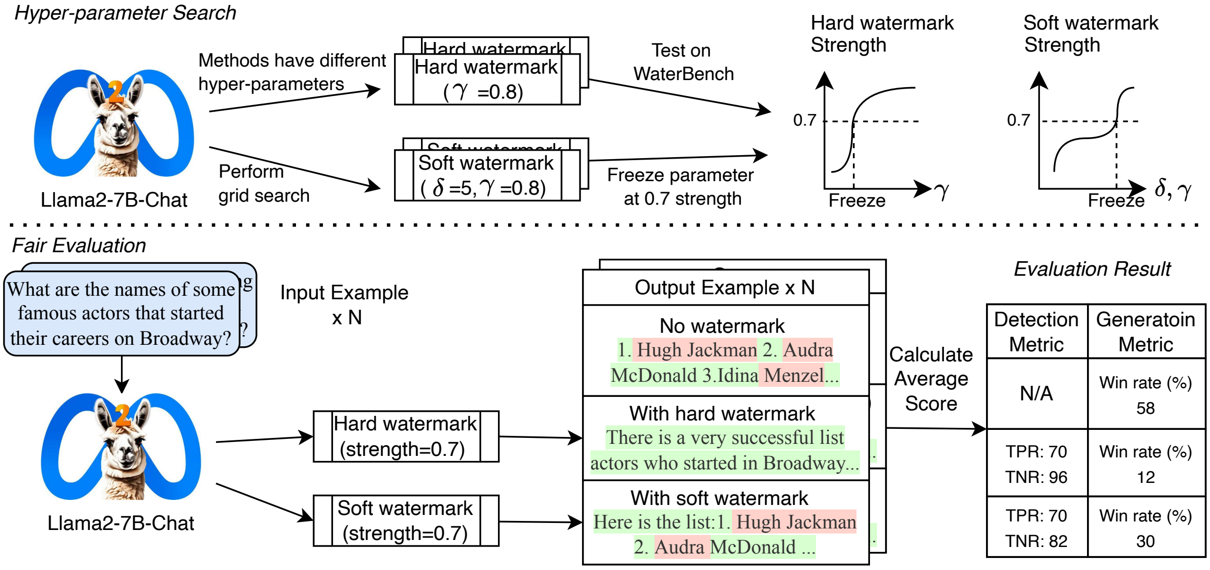


Figure 2: An illustration of the evaluation process on WaterBench. Given an LLM, a watermarking method and our benchmark, we first search the hyper-parameter to fix the watermarking strength of each method, then jointly evaluate their detection and generation performance for fair comparisons.

just the probability distribution at each decoding step, which guarantees the presence of detectable patterns, known as watermarks, in the generated text. Some works (Kirchenbauer et al., 2023b; Liu et al., 2023) focus on improving the detection robustness to paraphrasing attacks (Krishna et al., 2023) or at low-entropy environment (Lu et al., 2024). Other works like unbiased watermark (Hu et al., 2023) and NS-watermark (Takezawa et al., 2023) focus on improving the quality of generated texts (Hou et al., 2024; Li et al., 2023b).

On the other hand, post-hoc watermark (Atallah et al., 2001; Topkara et al., 2005) is also a line of research, which insert watermarks into texts by synonym replacement (Yang et al., 2023; Yoo et al., 2023) or paraphrasing (Munyer and Zhong, 2023). Recently, Sato et al. (2023) presented a simple but effective method that replace each space character with another codepoint of whitespace. However, this simple watermark can also be easily erased.

3 WaterBench

To investigate how inference-time watermarks perform on detection and generation, as shown in Figure 2, we propose a benchmarking procedure that ensures fair comparisons (Section 3.2). Then, we present the WaterBench dataset with a diverse length distribution (Section 3.3). Finally, we introduce the GPT4-Judge evaluation (Section 3.4).

3.1 Problem Definition for Watermarking

Generation Stage Assume an auto-regressive LLM θ has a vocabulary V, the probability distribution for the t-th token in a sequence $\mathbf{S} = \{s_1, s_2, \ldots, s_{|\mathbf{S}|}\}$ can be expressed as:

$$p(s_t) = p_\theta(s_t|s_{<t}) \tag{1}$$

LLM predicts the $p(s_t)$ by calculating a logit vector $l^{(t)} \in \mathbb{R}^{|V|}$ for each item k in vocabulary. Kirchenbauer et al. (2023a) propose 2 watermarks, namely hard and soft watermarks, to add watermarks to text by imposing restrictions on the vocabulary during each decoding step. Specifically, the "Hard Red List" watermark algorithm randomly divides the vocabulary into "green" and "red" lists using a hash function. During the generation process, only tokens from the green list can be chosen for the t-th position. While soft watermark approach introduces a constant δ to the logit $l_k^{(t)}$ of tokens in the green list during the prediction step:

$$p_k^{(t)} = \exp(l_k^{(t)} + \delta) / \sum_i \exp(l_i^{(t)}) \tag{2}$$

Detection Stage To detect the presence of the watermark in the generated text, statistical analysis techniques such as the *one proportion z-test* can be applied. While the hash function generates the green lists with a greenlist fraction γ, we can extract the watermark by re-computing the greenlist

Category & Source Data	ID	Task	Metric	Language	#data	Len.Input / Answer
(Short Input, Short Answer)						
KoLA (Yu et al., 2023)	1-1	Entity Probing	F1	English	200	7.72 / 2.96
Copen (Peng et al., 2022)	1-2	Concept Probing	F1	English	200	51.52 / 1.57
(Short Input, Long Answer)						
ELI5 (Fan et al., 2019)	2-1	Long-form QA	Rouge-L	English	200	41.04 / 236.6
FiQA (Maia et al., 2018)	2-2	Finance QA	Rouge-L	English	200	13.67 / 251.13
(Long Input, Short Answer)						
HotpotQA (Yang et al., 2018)	3-1	Multi-Doc QA	F1	English	200	10619.4 / 2.65
LCC (Chen et al., 2021)	3-2	Code Completion	Edit Sim	Python/C#/Java	200	2263.32 / 9.45
(Long Input, Long Answer)						
MultiNews (Fabbri et al., 2019)	4-1	Multi-Doc Summ.	Rouge-L	English	200	2198.65 / 260.88
QMsum (Zhong et al., 2021)	4-2	Query-Based Summ.	Rouge-L	English	200	12457.93 / 76.52
(Open Ended Generation)						
AlpacaFarm (Dubois et al., 2023)	5-1	Instruction Following	GPT4-Judge	English	805	32.58 / 64.13

Table 2: An overview of the dataset statistics in WaterBench. 'Dataset' denotes the origin of the sub-dataset. 'Len.Input / Answer' refer to the average length of input question and reference answer.

at each position to get a set of greenlist tokens $\mathbf{S}_g$. Then the significance can be derived by z-score:

$$z = (|\mathbf{S}_g| - \gamma|\mathbf{S}|)/\sqrt{\gamma(1-\gamma)|\mathbf{S}|} \qquad (3)$$

If the z-score is above the threshold, which means the corresponding P-value is small, then we can ensure that the text $\mathbf{S}$ is watermarked.

3.2 Benchmarking Procedure

Due to the two-stage nature of the task, previous works may use different hyper-parameters when testing the detection and generation, leading to unfair comparisons. As shown in Figure 2, we propose a fair benchmarking procedure to jointly evaluate the detection and generation performance.

Watermarking Strength. To retain consistency with the two stages and maintain fairness in our evaluations, we introduce the concept of watermarking strength (Mei et al., 2002) into LLM watermarks. In image watermarking (Akhaee et al., 2009), the higher watermarking strength means the better robustness for detecting the watermark. In LLM watermarking, we believe the watermarking strength should be independent with the referenced answer and can measure the detecting robustness. Therefore, we define the watermarking strength as the ratio of the number of watermarked texts that are correctly detected to the total number of watermarked texts, namely the TPR (True Positive Rate), which is a definite value after setting the input, the watermarking algorithm and its hyper-parameters. By freezing the watermarking strength, the evaluation results can remain consistent in the two stages. Some methods may adaptively set their strength for each case (Takezawa et al., 2023) and that's not discussed in this paper yet we leave for the future.

Hyper-Parameter Search. Although the watermarking strength depends on hyper-parameters, the hyper-parameters of different watermarking methods are not comparable (Ghosal et al., 2023). Therefore, we propose a hyper-parameter search procedure to unify the watermarking strength of different watermarking methods. Specifically, we first set the hyper-parameters of each watermarking method to the initial value by default, then we perform grid search (Alibrahim and Ludwig, 2021) to change the watermarking strength to the desired level and minimize the changes to hyper-parameters. Finally, we fix the hyper-parameters to the determined values and jointly evaluate the two-stage performance.

For researchers aiming to introduce a new watermark candidate to WaterBench, a suitable hyper-parameter should first be identified by them to achieve a certain True Positive Rate (TPR), for instance, 0.95. Subsequently, the evaluation code can be executed to obtain True Negative (TN) and Generation Metric (GM) results. Ultimately, they can benchmark their performance against other watermarks with an equivalent TPR on our benchmark.

3.3 Task Selection

As shown in Table 2, we select nine typical tasks for five distinct task settings, covering a wide range of input and output length, including:

Category 1: Short Input, Short Answer. As the input and answer length decides how much information that the watermarking algorithm can hide, we first choose two tasks that have short in-

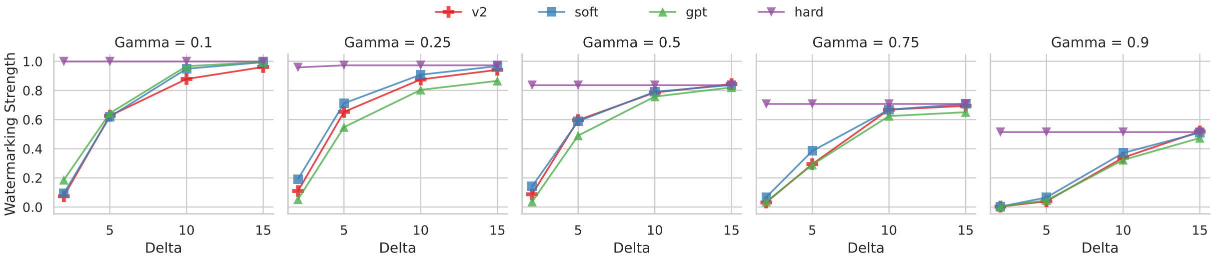


Figure 3: The watermarking strength results of 4 watermarking methods on Llama2-7B-chat after the hyper-parameter search for δ and γ. The watermarking strength is measured by the average TPR on our WaterBench.

put and answer length to disturb the watermarking methods. Both tasks evaluate the *Factual knowledge* in a close-ended setting. The task 1-1 is the knowledge probing, we use 200 triplets from KoLA dataset (Yu et al., 2023) with different frequency in Wikipedia to probe the facts from LLMs. For task 1-2, the concept probing, we use the 200 samples from the cic and csj task in Copen dataset (Peng et al., 2022). As the output length is short, we use the F1 score as the evaluation metric.

Category 2: Short Input, Long Answer. To control for the variable of answer length, we choose another 2 tasks with the short input but long answer. Both tasks belong to *Long-form QA*, which is the common format that users interact with LLMs, where user ask a short question and expect a long answer. For task 2-1, we use 200 samples from the ELI5 dataset, which is a long-form question-answering dataset composed of threads from the Reddit forum "Explain Like I'm five". For long-form QA with finance knowledge (Task 2-2), we use 200 samples from the FiQA dataset.

Category 3: Long Input, Short Answer. To control the variable of input length, we select 2 tasks from LongBench (Bai et al., 2023) with long input and short output. To evaluate the effect of watermarks of LLMs on the *reasoning* (Task 3-1), we select 200 samples from the HotpotQA dataset (Yang et al., 2018), which is a multi-hop question-answering dataset. For *code completion* (Task 3-2), we use 200 samples from the LCC dataset (Chen et al., 2021), a dataset constructed by filtering code within a single file from GitHub.

Category 4: Long Input, Long Answer. To control both input and output length, we involve 2 tasks with long inputs and answers. The 2 tasks are both *Summarization* task, which is a particular skill for serving people's information needs. We select 200 samples from the widely-used multi-document news summarization dataset, MultiNews (Fabbri et al., 2019). For query-based summarization, we use 200 samples from the QMSum dataset (Zhong et al., 2021) with both input documents and queries for specific parts of the documents.

Category 5: Open-Ended Generation. While the aforesaid datasets mainly evaluate the specific skills of LLMs, the input and output length may be limited to a certain range that is suited for the corresponding task. In real world application of LLMs, the abilities to follow the user's instructions are also important where the generation is often open-ended. To comprehensively evaluate the *instruction-following* performance of watermarked LLMs, we select the AlpacaFarm dataset (Dubois et al., 2023), which contains 805 instructions, consisting of 5 different sources of instructions, with 32.58 tokens in the input and 64.13 tokens in the reference answer on average.

3.4 Evaluation Metric

We used an evaluation method called GPT4-Judge (Zheng et al., 2023) to compare how well watermarked LLMs and Davinci-003 could generate text on the AlpacaFarm dataset. The GPT4-Judge measures which model's output the GPT-4 system prefers when shown two responses for the same instruction. To be fair, the order of the models' output texts are randomly mixed up before GPT-4 chose to avoid the position bias (Wang et al., 2023a).

4 Experiments

4.1 Experimental Settings

We choose 2 popular LLMs as our baselines: Llama2-7B-chat (Touvron et al., 2023) and Internlm-7B-8k (Team, 2023), both models are instruction-tuned to align with human preference. We evaluate 4 different representative watermarks on these 2 LLMs on our WaterBench, including:

Model	**C1: (Short Q, Short A)**				**C2: (Short Q, Long A)**				**C3: (Long Q, Short A)**			
	Factual Knowledge				*Long-form QA*				*Reasoning & Coding*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-7B-chat	–	–	17.8	–	–	–	21.3	–	–	–	37.5	–
+ hard watermark	89.5	100.0	5.0	↓71.9%	99.8	99.8	12.1	↓43.4%	82.5	100.0	16.4	↓56.3%
+ soft watermark	90.2	98.8	7.7	↓56.6%	100.0	100.0	9.9	↓53.3%	80.2	100.0	19.8	↓47.2%
+ gpt watermark	95.8	100.0	13.6	↓23.9%	100.0	92.5	5.2	↓75.5%	85.0	98.9	14.7	↓60.7%
+ v2 watermark	83.8	100.0	11.2	↓37.1%	100.0	100.0	13.3	↓37.4%	81.0	100.0	13.9	↓63.0%
Internlm-7B-8k	–	–	26.3	–	12.2	–	18.4	–	–	–	31.6	–
+ hard watermark	88.2	99.6	1.8	↓93.1%	96.0	99.8	9.6	↓47.9%	88.5	99.7	11.7	↓63.0%
+ soft watermark	80.7	99.6	6.2	↓76.3%	99.7	99.8	7.6	↓58.7%	89.8	99.7	10.6	↓66.5%
+ gpt watermark	97.8	100.0	3.2	↓87.8%	100.0	100.0	7.8	↓57.6%	86.0	100.0	11.4	↓63.8%
+ v2 watermark	85.6	100.0	11.0	↓58.4%	98.0	100.0	7.6	↓58.4%	92.6	100.0	15.8	↓50.1%

Table 3: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for category 1, 2 and 3 tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

Model	**C4: (Long Q, Long A)**				**C5: Open-Ended**				**Overall: (12345)**			
	Summarization				*Instruction Following*				*Detection & Generation*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-7B-chat	–	–	23.3	–	–	–	54.7	–	–	–	28.3	–
+ hard watermark	100.0	100.0	11.6	↓50.0%	100.0	98.8	1.1	↓98.0%	95.3	99.5	10.1	↓64.1%
+ soft watermark	100.0	100.0	10.2	↓56.3%	99.4	98.9	0.6	↓98.9%	94.9	99.5	10.7	↓62.3%
+ gpt watermark	100.0	99.8	7.2	↓69.1%	99.6	95.8	0.2	↓99.5%	96.7	96.9	9.1	↓67.9%
+ v2 watermark	100.0	99.8	11.6	↓50.2%	100.0	99.9	0.9	↓98.4%	94.1	99.9	11.2	↓60.3%
Internlm-7B-8k	–	–	17.8	–	–	–	21.5	–	–	–	23.3	–
+ hard watermark	96.0	100.0	6.4	↓64.3%	96.5	99.6	0.8	↓96.5%	93.7	99.7	6.6	↓71.6%
+ soft watermark	98.7	99.8	4.6	↓74.0%	97.4	99.5	0.3	↓98.6%	94.0	99.6	6.5	↓72.2%
+ gpt watermark	97.2	100.0	5.2	↓70.9%	99.3	99.4	0.5	↓97.7%	96.8	99.8	6.2	↓73.4%
+ v2 watermark	97.2	100.0	5.5	↓69.2%	97.7	99.4	0.5	↓97.7%	95.1	99.8	8.9	↓61.7%

Table 4: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for category 4, 5 and all tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

- **Hard Watermark**: The hard watermark (Kirchenbauer et al., 2023a) is a binary watermark that restricts the vocabulary of the model to a subset of words during decoding.
- **Soft Watermark**: The soft watermark (Kirchenbauer et al., 2023a) is a continuous watermark that divide γ vocabulary and adds a constant δ on logits to encourage watermarked vocabularies.
- **GPT Watermark**: The GPT watermark (Zhao et al., 2023) simplifies the watermarking process with a fixed group of restricted vocabularies to improve robustness against editing attacks.
- **V2 Watermark**: The V2 watermark (Kirchenbauer et al., 2023b) improves the soft watermark with different hashing schemes, including the LeftHash and SelfHash to secure better robustness to the paraphrasing attack.

As the evaluation procedure described in Section 3.2, we first adjust the watermarking strength of each watermark by grid search. As shown in Figure 3, we find that the watermarking strength of each watermark increases when δ increases and increases when γ reduces. We then choose the watermarking strength of 0.95 and 0.7 for each watermark and freeze their hyper-parameters for further detection and generation evaluation. Apart from the grid search results, we also display the ROC curve of watermarks in Appendix A.2.

4.2 Main Results

We conduct evaluations at the 0.95 watermarking strength on each task and report watermarks' results in Table 3 and 4. Here are our findings:

Detection Performance. Among all tasks, the detection performance for the short answer tasks (Category 1 and 3) are significantly worse than other tasks. This is because that watermarked LLMs produce short responses for these tasks, which can not contain enough green-words for detection, re-

Model	**C1: (Short Q, Short A)**				**C2: (Short Q, Long A)**				**C3: (Long Q, Short A)**			
	Factual Knowledge				*Long-form QA*				*Reasoning & Coding*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-7B-chat	–	–	17.8	–	–	–	21.3	–	–	–	37.5	–
+ hard watermark	0.0	100.0	13.7	↓ 23.3%	100.0	100.0	19.4	↓ 8.9%	39.2	100.0	21.0	↓ 44.1%
+ soft watermark	0.0	100.0	13.8	↓ 22.6%	100.0	100.0	19.4	↓ 8.7%	41.2	100.0	20.6	↓ 45.1%
+ gpt watermark	11.8	100.0	17.0	↓ 4.4%	99.5	99.8	13.8	↓ 35.0%	25.1	100.0	17.3	↓ 53.9%
+ v2 watermark	0.0	100.0	14.9	↓ 16.6%	99.5	100.0	19.4	↓ 8.8%	39.8	100.0	25.1	↓ 33.2%

Model	**C4: (Long Q, Long A)**				**C5: Open-Ended**				**Overall: (12345)**			
	Summarization				*Instruction Following*				*Detection & Generation*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-7B-chat	–	–	23.3	–	–	–	54.7	–	–	–	28.3	–
+ hard watermark	91.8	100.0	19.9	↓ 14.4%	96.5	99.8	17.3	↓ 68.4%	70.7	99.9	18.4	↓ 35.1%
+ soft watermark	92.0	100.0	20.2	↓ 13.3%	95.4	99.8	19.0	↓ 65.2%	70.7	99.9	18.6	↓ 34.4%
+ gpt watermark	96.0	100.0	15.0	↓ 35.4%	93.4	99.9	4.1	↓ 92.5%	69.9	99.9	14.5	↓ 48.7%
+ v2 watermark	88.8	100.0	19.7	↓ 15.3%	94.0	99.9	17.0	↓ 68.9%	69.4	100.0	19.5	↓ 31.2%

Table 5: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for all tasks at the watermarking strength level of 0.7 with z-score threshold of 4 for Llama2-7B-chat.

sulting in low z-scores, making the detectors more likely to fail on discovering the watermark.

For the overall detection performance, most watermarks can achieve high TP rates of around 95% which is consistent with the fixed watermarking strength, while the TN rates are nearly 100% for all methods. This indicates that current LLM watermarks (Kirchenbauer et al., 2023a) are generally good at detecting watermarked texts while remaining a clear distinction from unwatermarked texts.

Generation Performance. However, in terms of generation quality, all watermarks lead to significant drops in GM compared to the original models. The hard watermark exhibits the largest decreases of over 50% in most cases. The generation performance also declines more severely for the open-ended task with over 90% drop. These findings suggest that current watermarks encounter challenges in maintaining the generation quality, particularly for instruction-following tasks.

Among different watermarks, as shown in Table 4, the V2 watermark achieves higher GMs in most task categories, highlighting its effectiveness in preserving generation quality. The soft watermark and GPT watermark also exhibit competitive performance. While V2 watermark even shows better True Negative rate in most categories, indicating its advantage on watermark detection, too.

In addition, we observe larger performance drops for InternLM compared to Llama2 under the same watermarking method, indicating that impact of watermarking can vary among LLMs, highlighting the importance of model-specific evaluations.

In summary, while current watermarks are effective in detection, their generation quality still degrades significantly. Future work can explore new watermark designs to minimize such declines.

4.3 Watermarking Strength Analysis

To analyze the influence of the watermarking strength on evaluating the detection performance and generation quality, we conduct experiments for the 4 watermarking methods with 0.7 watermarking strength at Table 5 to compare with the main results in Section 4.2 at 0.95 watermarking strength. And we have the following observations:
(1) There exists a trade-off between the watermarking strength and generation quality. Models tend to exhibit larger drops in GM at 0.95 strength compared to 0.7. For example, the watermark with the worst generation score in Table 5 (0.7 strength) can rank the first in Table 4 (0.95 strength), which can't reflect the real difference between watermark algorithms. This highlights the importance of using a standardized strength for fair comparisons.
(2) At a lower strength of 0.7 , average TP rates drop noticeably compared to 0.95 strength across different tasks. We observe the largest TP rate drop (from ∼90% to ∼0%) in Category1 with short input and short answer. This suggests that our WaterBench is hard enough to add strong disturbance on watermarks for adjusting watermarking strengths.
(3) V2 watermark maintains relatively stable detection and generation performance at both strengths, outperforming other methods. However, V2 water-

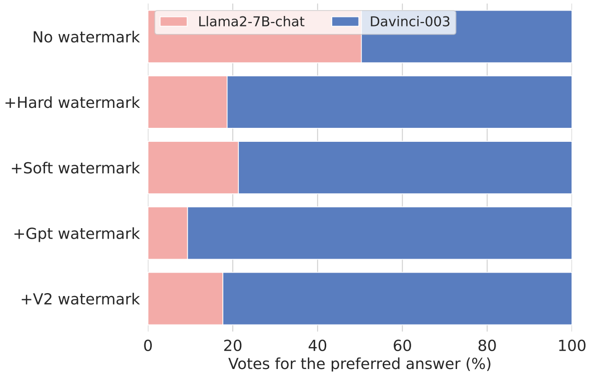


Figure 4: Average votes by three human annotators for the preferred answer between our watermarked LLM generation and text-davinci-003 baseline response.

mark still makes Llama2's generation performance drops 31.2%, indicating that further exploration is needed to minimize quality degradation.

4.4 Human Evaluation

To prove the effectiveness of GPT4-Judge on task 5-1, we conduct a human evaluation that annotate the actual human preferences on model responses. We sample 100 generation results respectively from the 5 models at the watermarking strength level of 0.7. Then we ask three human annotators to vote for their preferred response between the watermarked LLM and Davinci-003 baseline (See Appendix A.3 for more details). In total, we collect 1, 500 human feedback, yielding the following findings:

(1) The results from the three human annotators align with the labeling from GPT4. Figure 4 exhibits the average voting results of three humans for the instruction-following task, where Llama2-7B-Chat without watermark achieves a 50.3% win rate against the Davinci-003 baseline, which is consistent with the GPT4's simulated win rate of 54.7%. Besides, the other watermarked LLMs also obtain the similar win rates to GPT4's predictions, further demonstrating the effectiveness of GPT4-Judge.

(2) The inter-annotator agreement coefficients between GPT4 and three human annotators are varied, but all of them are above 0.6, indicating substantial agreement. As shown in Figure 5, GPT4 has a 0.83 agreement with human1, which can be viewed as almost perfect agreement. While GPT4 only gets substantial agreement with human2 and human3. Additionally, the agreements among three human annotators are around 0.6, which means substantial agreement. So there also exists a variety of human annotators, which may lead to the different agreements with GPT4-Judge (Dubois et al., 2023).

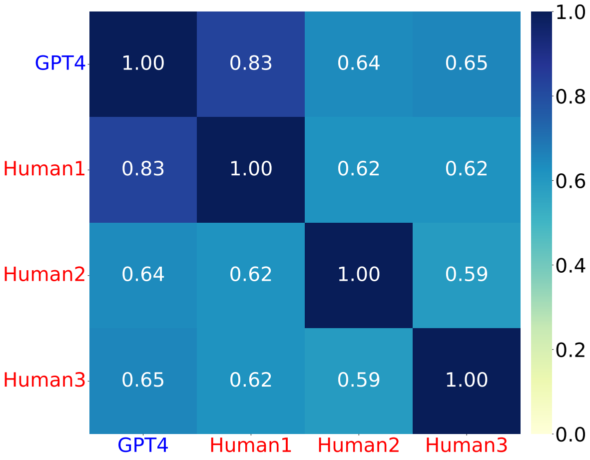


Figure 5: Cohen's kappa coefficient for inter-annotator agreement among GPT4 and human annotators.

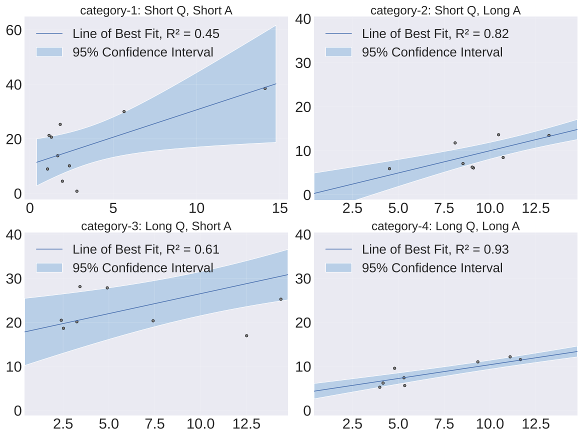


Figure 6: Scatter plots for each pair of tasks (e.g. 1-1 and 1-2), each point is an evaluated model's GM scores of two tasks in the same category.

4.5 Correlation Analysis

To verify the diversity of our task selection, we analyze the inner task performance correlation for categories. As plotted in Figure 6, the generation performances of the watermarked LLMs on two sub-tasks of each category reveal a clear linear correlation, indicating the reliability of our task categorization. Notably, there is a more pronounced performance gap between tasks in the shorter answer categories (1 and 3). The convergence of task performances may reflect the model's generalization capability on different tasks. Overall, WaterBench provides a comprehensive and challenging benchmark for evaluating LLM watermarks.

5 Conclusion

In this paper, we propose WaterBench, a new benchmark for evaluating large language model watermarks. We first introduce a benchmarking procedure that searches hyperparameters to ensure consistent watermarking strength across different methods, allowing for a fair comparison. Second, we construct a multi-task benchmark spanning nine typical NLP tasks with varying input/output lengths. Finally, we incorporate the GPT4-Judge metric to automatically evaluate the results. Experiments show that it sensitively reflects declines in instruction-following quality after watermarking. We hope that our work will inspire and facilitate the future research on LLM watermarks.

Acknowledgement

This work is supported by a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003), Tsinghua University Initiative Scientific Research Program and Zhipu AI.

Limitations

Although we have conducted extensive experiments, there are still some limitations for our work: (1) The detection candidate is only the reference answer in the benchmark, which is mainly written in the human expert style (Ghosal et al., 2023). However, all texts without the watermark can be considered as negative examples. (2) There is only one generation metric for each task. We will explore more metrics such as BertScore (Zhang et al., 2019) and FactCC (Kryściński et al., 2019) to evaluate the performance of LLMs in different aspects. (3) A watermarking method may have different compositions for hyper-parameters to achieve the same watermarking strength, while in our experiments we only evaluate one composition with the minimum changes to hyper-parameters. We encourage the future research to explore this composition.

Ethics Statement

In this section, we will discuss the ethical consideration for our work.

Licenses. For open-accessible datasets used in our work, we have checked their licenses. The KoLA (Yu et al., 2023) dataset is shared under the GPLv3 license, Copen (Peng et al., 2022) is shared under the MIT license, ELI5 (Fan et al., 2019) is shared under the BSD license, the LongBench (Bai et al., 2023) which includes task 3-1 to 4-2 is released under the MIT license, and the AlpacaFarm dataset (Dubois et al., 2023) is shared under the Apache-2.0 license. The Licenses for the large language models are also available. Llama2-7B-chat (Touvron et al., 2023) is released under the Meta License which needs to apply on their websites, and InternLM-7B-8k (Team, 2023) is shared under the Apache-2.0 license.

Ethics Considerations for AI assistants AI assistants like GPT4 are powerful, even our automatic evaluation process has adapted GPT4 as an evaluator, which complies with the AI ethical guidelines set by the European Union[1]. These guidelines place emphasis on various ethical aspects, including technical robustness, safety, privacy, transparency, and accountability. We make sure that the usage of AI systems in our research are aligned with these principles. They also highlight the importance of ensuring the safety of AI systems and establishing accountability mechanisms for potential negative consequences. This encourages our work to evaluate LLM watermarks that may help policy makers conduct regulations for generative AI systems with detectable watermarks.

References

Mohammad Ali Akhaee, S Mohammad Ebrahim Sahraeian, Bulent Sankur, and Farokh Marvasti. 2009. Robust scaling-based image watermarking using maximum-likelihood decoder with optimum strength factor. *IEEE Transactions on Multimedia*, 11(5):822–833.

Hussain Alibrahim and Simone A Ludwig. 2021. Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1551–1559. IEEE.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. 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.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang,

---

[1] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`

and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. 2023. Does chatgpt resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of ACL*, pages 3558–3567.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. *arXiv preprint arXiv:2308.00113*.

Yu Fu, Deyi Xiong, and Yue Dong. 2023. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. *arXiv preprint arXiv:2307.13808*.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Abe Bohan Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. 2024. k-semstamp: A clustering-based semantic watermark for detection of machine-generated text.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*.

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

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2023b. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Yuhang Li, Yihan Wang, Zhouxing Shi, and Cho-Jui Hsieh. 2023b. Improving the generation quality of watermarked large language models via word importance scoring.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2023. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*.

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

Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024. An entropy-based text watermarking detection method.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Shi-chun Mei, Ren-hou Li, Hong-mei Dang, and Yun-kuan Wang. 2002. Decision of image watermarking strength based on artificial neural-networks. In *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, volume 5, pages 2430–2434. IEEE.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Travis Munyer and Xin Zhong. 2023. Deeptextmark: Deep learning based text watermarking for detection of large language model generated text. *arXiv preprint arXiv:2305.05773*.

OpenAI. 2023. Gpt-4 technical report,. *OpenAI*.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, and Irwin King. 2024. Markllm: An open-source toolkit for llm watermarking.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*.

Yuki Takezawa, Ryoma Sato, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Necessary and sufficient watermark for large language models. *arXiv preprint arXiv:2310.00833*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

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

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

Shangqing Tu, Chunyang Li, Jifan Yu, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2023. Chatlog: Recording and analyzing chatgpt across time. *arXiv preprint arXiv:2304.14106*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023b. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. *arXiv preprint arXiv:2306.07401*.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. 2023. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust natural language watermarking through invariant features. *arXiv preprint arXiv:2305.01904*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

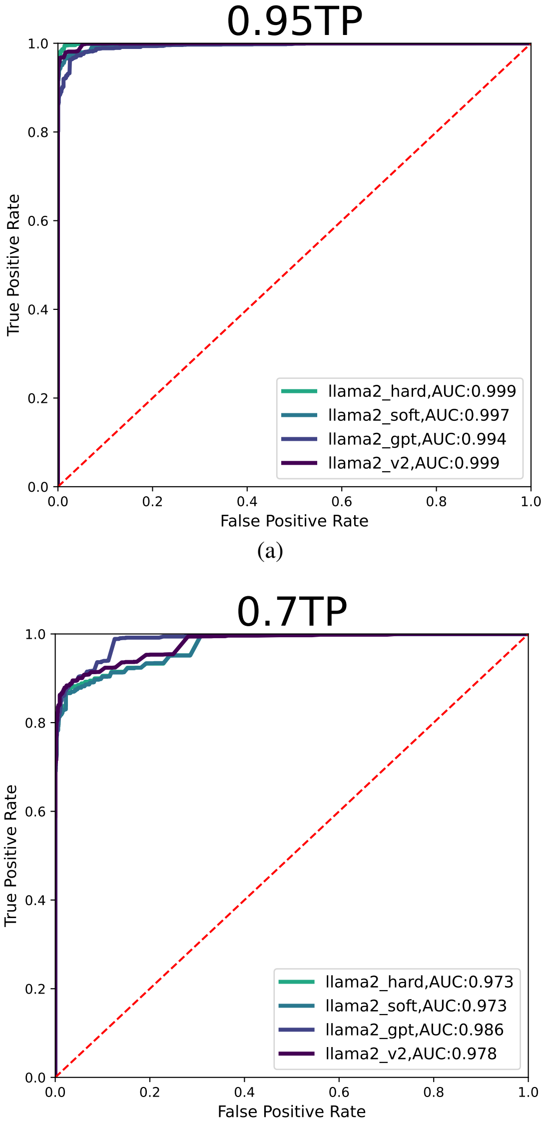
A Implementation Details

A.1 Deployment Details

In our evaluating and detecting experiments, we utilize the widely-used *Pytorch* and *transformers* library to load all the models. All the experiments are conducted on Ubuntu 20.04.4 server equipped with 112 Intel Xeon(R) Platinum 8336C CPU cores, and graphic cards that contained 8 NVIDIA A100 SXM 80GB GPUs. Besides, the CUDA version is 11.4, the Python version is 3.10.11, the PyTorch version is 2.0.1 and the transformers version is 4.31.0. We integrate the code from LM-Watermark[2], V2 Watermark[3] and GPT Watermark[4] to implement a unified watermarking experiment tool, where different kinds of watermark can be evaluated fairly. The code of our all-in-one tool is provided in the supplement files.

A.2 Hyper-parameters Search Details

In order to obtain experimental groups with the same watermark strength, there are three hyper-parameters that need to be obtained through search. The first is the vocabulary partition parameter γ, which represents the proportion of the green list vocabulary within the model's total vocabulary. The second is the bias constant δ for the logit, representing the hardness of the red list vocabulary. And the last one is the threshold used in the Z-test. Under the same threshold conditions, according to the calculation method of the z-score, it can be observed in Figure 3 that when γ increases, the average z-score will decrease, resulting in a weaker watermark strength. And the increase in δ implies a stronger hardness of the watermark, which in turn results in a stronger watermark strength. Therefore, we first set the same threshold and adjust these two hyper-parameters, γ and δ, based on their relationship with watermark strength to find different watermark groups with the same strength category. Then, we make slight adjustments to the threshold for these watermark groups to ensure they achieve the same strength level with greater precision. Using this approach, we obtained the appropriate hyper-parameters and recorded them in Table 6. Note that the default z-score threshold is 4, which is a commonly used value in prior works (Kirchenbauer et al., 2023a,b).



(a)

(b)

Figure 7: ROC curves for two watermark strengths.

To prove the effectiveness of our hyper-parameter searching process, we draw the ROC curve in Figure 7 by adjusting the z-score threshold. First, all of the 4 kinds of watermarks on the same watermark level obtain similar AUC scores, where even all AUC socres are 0.99 at the 0.95

---

[2] `https://github.com/jwkirchenbauer/lm-watermarking`

[3] `https://github.com/jwkirchenbauer/lm-watermarking/tree/main/watermark_reliability_release`

[4] `https://github.com/XuandongZhao/GPTWatermark`

watermark strength. As there are only small difference among watermarks on the ROC curve, it is reasonable for us to set the same initial z-score threshold for all models. Second, when at the perfect point that when False Positive Rate is 0, the True Positive Rate is often not at the target level for some watermarks. That's why we need to adjust the z-score threshold to reach the watermark strength in the end. Finally, as we first adjust the γ and δ to get different watermark strengths, we can observe the difference of 0.7 and 0.95 strengths caused by these hyper-parameters on the sub-figures in Figure 7. If we fix the γ and δ, and only adjust the z-score threshold, then we may not get the ideal TPR that can be reached by adjusting γ and δ.

Module	**Parameter**	**Value**
Llama2 + hard watermark	γ 0.95TP z-score threshold	0.25 4.3
Llama2 + soft watermark	γ δ 0.95TP z-score threshold	0.1 10 4.0
Llama2 + gpt watermark	γ δ 0.95TP z-score threshold	0.1 10 4.2
Llama2 + v2 watermark	γ δ 0.95TP z-score threshold	0.25 15 4.1
Llama2 + hard watermark	γ 0.7TP z-score threshold	0.75 4.2
Llama2 + soft watermark	γ δ 0.7TP z-score threshold	0.75 15 4.2
Llama2 + gpt watermark	γ δ 0.7TP z-score threshold	0.65 12.5 4.0
Llama2 + v2 watermark	γ δ 0.7TP z-score threshold	0.75 15 3.8
Internlm + hard watermark	γ 0.95TP z-score threshold	0.15 3.5
Internlm + soft watermark	γ δ 0.95TP z-score threshold	0.1 10 3.2
Internlm + gpt watermark	γ δ 0.95TP z-score threshold	0.25 15 4.1
Internlm + v2 watermark	γ δ 0.95TP z-score threshold	0.1 10 4.0

Table 6: Hyper-parameters for each model.

It is noted that there may exist many points that satisfy the fixed TPR, or quite close to the value, to select hyper-parameters, we adopt the following approach:

First, we use a grid search approach to find the proper points. As shown in Table 7 and Table 8, there may be many points proximate to a TPR of 0.95, albeit not precisely equal to it. We choose the points that exhibit the least deviation, like TPR=0.949 to report.

Subsequently, we examine two soft watermark results that approximated a TPR of 0.95, as derived from the hyperparameter search. For the overall scores, both two watermarks are around the level of TPR=0.95. Their TNR scores are even the same as 0.995 while their GM scores are a little different, one is 10.7 and the other is 11. Despite the disparity in their GM scores on C1 and C2, the scores on C3 and C4 are akin. Therefore, we generally ensure minimal differences exist between points around a TPR of 0.95. We choose to report the TPR=0.949 one over TPR=0.967.

We acknowledge that a more comprehensive comparison would involve analyzing the Pareto frontier of "TPR @ fixed TNR" versus "GM". This would provide a more accurate assessment of the trade-offs between these metrics across the different methods. However, given the need for extensibility in adding new watermarks and the computational cost of GPU resources, our current evaluation framework may not be able to accommodate the Pareto frontier analysis.

A.3 Human Annotation Details

To investigate human preferences for the results of task 5-1, we recruited three human annotators from three prominent universities in our country. Among them, two annotators are male and one is female. All participants hold at least a bachelor's degree.

We have established working contracts with all three annotators, ensuring compensation in accordance with mutually agreed-upon wage standards and working hours. These employment arrangements are in compliance with the local regulations.

The annotation instructions are presented in Table 9. To develop a suitable protocol for our task, we consulted relevant prior works (Dubois et al., 2023; Zheng et al., 2023). Moreover, we subjected this data collection protocol to review by two PhD students to mitigate potential ethical risks.

Model	C1: (Short Q, Short A)				C2: (Short Q, Long A)				C3: (Long Q, Short A)			
	Factual Knowledge				*Long-form QA*				*Reasoning & Coding*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-7b-chat	–	–	17.8	–	–	–	21.3	–	–	–	37.5	–
+ soft watermark γ=0.1, δ=10	90.2	99.2	7.7	↓ 56.6%	100.0	100.0	9.9	↓ 53.3%	80.2	100.0	19.8	↓ 47.2%
+ soft watermark γ=0.25, δ=15	95.5	100.0	4.7	↓ 73.4%	100.0	99.8	13.0	↓ 38.7%	84.8	100.0	19.3	↓ 48.6%

Table 7: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for category 4, 5 and all tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

Model	C4: (Long Q, Long A)				C5: Open-Ended				Overall: (12345)			
	Summarization				*Instruction Following*				*Detection & Generation*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-7b-chat	–	–	23.3	–	–	–	54.7	–	–	–	28.3	–
+ soft watermark γ=0.1, δ=10	100.0	100.0	10.2	↓ 56.3%	99.4	98.9	0.6	↓ 98.9%	94.9	99.5	10.7	↓ 62.3%
+ soft watermark γ=0.25, δ=15	100.0	100.0	12.2	↓ 47.6%	99.9	98.8	0.6	↓ 98.9%	96.7	99.5	11.0	↓ 61.0%

Table 8: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for category 4, 5 and all tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

INSTRUCTION: In this task, we will ask you to select the preferred output AI model's responses to instructions.
You will read a batch of examples, which are composed of the following:
1. an Instruction we give to the AI system
2. an Input that is provided along with the instruction
3. Output (a), the first output from the AI system
4. Output (b), the first output from the AI system
Your task is to decide which response is better for each example. There are several dimensions that you can think along. Consider the following questions:
1. Is the response helpful? For example, if the instruction asked for a recipe for healthy food, and the response is a useful recipe, then we can consider it helpful.
2. Is the response language natural? For example, AI responses often have repetitions, which is not natural.
3. Is the response factual/accurate? For example, AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate.
4. and so on ... ultimately, you should decide which response is better based on your judgment and based on your own preference.
You should answer using only Output (a) or Output (b) depending on which response is better.

Table 9: Instruction for human annotators.

B Evaluation Details

B.1 Full Results

In section 4.2, we introduce the average generation results of each layer. Due to the page limit, the detailed evaluation results of each sub-task are not fully presented. In this section, we report the full evaluation results for all tasks. As shown in Table 10 and 11 , from category 1 to category 4, each category has 2 sub-tasks with similar input and answer length. For category 2 and 4, the generation metric scores of sub-tasks in each layer are in the similar range, while sub-tasks in category 1 or 3 are not similar. For example, Llama2-7B-chat achieves 30.0 on task 1-2 but only 5.7 on task 1-1 although 1-1 and 1-2 are classified into the same category. This difference of tasks in same category exhibits the task diversity of our benchmark, showing that even if two tasks have similar input and output length range, models can get different scores on them, which proves the necessity of using 2 tasks for each category to test multiple aspects of LLM's ability.

B.2 Case Study

This section contains sampled examples from every evaluation task of the WaterBench. The following tables from Table 19 to Table 27 show six different answers to the same question, including human answer, the answer of Llama2-7B-Chat without watermark, and the answers of Llama2-7B-Chat with four kinds of watermarks at the same 0.95 watermarking strength. By observing these real responses, we have some interesting findings:

(1) **Bypass safety constrains**: Sometimes Llama-7B-Chat refuses to answering some risky questions, while after adding the watermarking algorithms, the LLM can give answers to these questions. For example, in Table 19, LLM without watermark refuse to answering the personal information of a person entity on WikiPedia, but the LLM with the GPT Watermark mentions that the person is from Hawaii according to the public information. This may be because of the biased decoding process of watermarking can bypass some constrains that LLMs have learned in safety alignment.

(2) **Repeating sequences**: As shown in Table 22, both the soft and GPT watermark produces repeating words like "- In recent history" or "Ghana supplies". Besides, this kind of repeating usually won't end until the generated text length reaches the max length limit. This never-ending generation process may be because of the lowered probability on <eos> token while the tokens in the repeating sequences are allocated with higher probabilities. Some watermarks (Kirchenbauer et al., 2023a) use the hashing mechanism that depends on the short context to decide the green list. Therefore, the repeating tokens may create a loop for the green lists, where the repeating tokens are favored by the green list and the context consisting of repeating tokens are hashed into the random number that produces the same green lists.

(3) **Symbol Replacement**: For the instruction-following AloacaFarm dataset (Dubois et al., 2023), there are some instructions that require listing some points, which are usually organized with '•' in the markdown format. As shown in Table 27, however, the hard watermark produces '-' and the v2 watermark produces '*', which is a rarely used symbol for listing. We assume that the biased distribution of tokens may forbid some common symbols, which lead to the changes in symbols. This phenomenon suggests that the watermarking process may introduce changes to the content and style of LLM-generated responses. Understanding these changes and their potential implications is crucial for evaluating the performance of watermarked LLMs.

B.3 Results on Another LLM

In order to show the generalizability of our benchmark across the broad spectrum of existing and future LLMs as well as to demonstrate the scalability of our benchmark to model size, We evaluate one more popular LLM(Llama2-13B-chat) in the experiments. The result is shown in Table 13 and Table 14.

It is noted that all watermarks on Llama2-13B-chat exhibit the greatest performance drop on instruction-following tasks (C5) among all tasks, which is consistent with the observation on Llama2-7B-chat in Section 4.

B.4 Results on Another Watermark

In addition to the LLM watermarks described in the previous section, we also evaluate the performance of our benchmark on another unbiased watermarking scheme using Gumbel tricks (Hu et al., 2023).

Model	Category1: (Short Input, Short Answer)						Category2: (Short Input, Long Answer)					
	1-1			1-2			2-1			2-2		
	TP	**TN**	**GM**	**TP**	**TN**	**GM**	**TP**	**TN**	**GM**	**TP**	**TN**	**GM**
Llama2-7B-chat	–	–	5.7	–	–	30.0	–	–	21.3	–	–	21.3
+ hard watermark	100.0	100.0	1.1	79.0	100.0	8.9	100.0	99.5	10.5	99.5	100.0	13.6
+ soft watermark	97.5	99.3	1.7	82.9	98.0	13.8	100.0	100.0	8.1	100.0	100.0	11.8
+ gpt watermark	96.0	100.0	1.8	95.5	100.0	25.3	100.0	91.5	4.5	100.0	93.5	5.9
+ v2 watermark	100.0	100.0	1.1	67.5	100.0	21.3	100.0	100.0	13.2	100.0	100.0	13.5
Internlm-7B-8k	–	–	14.1	–	–	38.5	–	–	17.9	–	–	18.9
+ hard watermark	85.9	99.4	2.8	90.8	100.0	0.8	94.0	99.5	10.7	98.0	100.0	8.4
+ soft watermark	93.4	99.4	2.4	68.0	100.0	10.1	99.5	100.0	9.1	100.0	99.5	6.1
+ gpt watermark	98.0	100.0	1.9	97.5	100.0	4.5	100.0	100.0	8.5	100.0	100.0	7.1
+ v2 watermark	91.4	100.0	1.3	76.9	100.0	20.6	98.0	100.0	9.0	98.0	100.0	6.3

Table 10: True Positive Rate (TP), True Negative Rate (TN) and Generation Metric (GM) for category 1 and 2 tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

Model	Category3: (Long Input, Short Answer)						Category4: (Long Input, Long Answer)					
	3-1			3-2			4-1			4-2		
	TP	**TN**	**GM**	**TP**	**TN**	**GM**	**TP**	**TN**	**GM**	**TP**	**TN**	**GM**
Llama2-7B-chat	–	–	25.0	–	–	50.0	–	–	25.9	–	–	20.7
+ hard watermark	72.0	100.0	4.9	93.0	100.0	27.8	100.0	100.0	11.1	100.0	100.0	12.2
+ soft watermark	62.0	100.0	14.4	98.5	100.0	25.3	100.0	100.0	9.3	100.0	100.0	11.0
+ gpt watermark	70.0	98.2	12.5	100.0	99.5	17.0	100.0	100.0	4.8	100.0	99.5	9.6
+ v2 watermark	67.3	100.0	7.4	94.5	100.0	20.4	100.0	99.5	11.7	100.0	100.0	11.5
Internlm-7B-8k	–	–	25.0	–	–	38.2	–	–	20.2	–	–	15.4
+ hard watermark	83.3	99.4	3.2	93.4	100.0	20.1	96.0	100.0	5.3	96.0	100.0	7.4
+ soft watermark	84.5	99.4	2.5	94.5	100.0	18.6	99.5	99.5	4.0	97.9	100.0	5.3
+ gpt watermark	85.6	100.0	2.4	86.3	100.0	20.5	99.5	100.0	4.2	94.9	100.0	6.2
+ v2 watermark	87.7	100.0	3.4	97.5	100.0	28.1	98.5	100.0	5.3	96.0	100.0	5.6

Table 11: True Positive Rate (TP), True Negative Rate (TN) and Generation Metric (GM) for category 3 and 4 tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

Model	5-1: Open-Ended			Overall		
	TP	**TN**	**GM**	**TP**	**TN**	**GM**
Llama2-7B-chat	–	–	54.7	–	–	28.3
+ hard	100.0	98.8	1.1	95.3	99.5	10.1
+ soft	99.4	98.9	0.6	94.9	99.5	10.7
+ gpt	99.6	95.8	0.2	96.7	96.9	9.1
+ v2	100.0	99.9	0.9	94.1	99.9	11.2
Internlm-7B-8k	–	–	21.5	–	–	23.3
+ hard	96.5	99.6	0.8	93.7	99.7	6.6
+ soft	97.4	99.5	0.3	94.0	99.6	6.5
+ gpt	99.3	99.4	0.5	96.8	99.8	6.2
+ v2	97.7	99.4	0.5	95.1	99.8	8.9

Table 12: True Positive Rate (TP), True Negative Rate (TN) and Generation Metric (GM) for Open-ended generation and all tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

We conduct evaluations to the watermarks on each task with Llama2-7b-chat and report watermarks' results in Table 15 and Table 16. Note that We evaluate the two schemes when LLR(Log likelihood ratio) score threshold of the whole sentence is 10, which means a p-value of less than 0.0005 is ensured.

We find that although the GMs of the unbiased watermark are quite high, the TP rates are not as satisfactory. This result demonstrates the trade-off between the watermarking strength and generation quality.

B.5 Watermark Computational Efficiency

In addition to the performance of watermarking methods, we also evaluate the average decoding speed of different watermarking methods comprehensively.

As the results shown in the Table 17, the differences between watermark schemes don't have a large impact on the computational efficiency during the model inference.

B.6 Standard Deviation for GMs Scores

Using the results from our prior experiments conducted on Llama2-7B-chat at a True Positive Rate (TPR) of 0.95, we have calculated the average and standard deviation for the GM scores.

As the results shown in the Table 18, the standard deviation for the GM scores is relatively small

Model	C1: (Short Q, Short A) *Factual Knowledge*				C2: (Short Q, Long A) *Long-form QA*				C3: (Long Q, Short A) *Reasoning & Coding*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-13B-chat	–	–	10.5	–	–	–	22.2	–	–	–	29.2	–
+ hard watermark	97.0	100.0	3.1	↓ 70.7%	100.0	100.0	15.8	↓ 28.8%	90.2	100.0	16.7	↓ 42.9%
+ soft watermark	85.0	100.0	2.1	↓ 79.7%	100.0	99.8	15.4	↓ 30.4%	94.0	100.0	16.2	↓ 44.4%
+ gpt watermark	90.0	100.0	5.9	↓ 43.8%	99.5	92.5	5.0	↓ 77.6%	89.2	99.0	11.6	↓ 60.1%
+ v2 watermark	74.5	100.0	10.0	↓ 5.3%	99.5	100.0	13.9	↓ 37.3%	89.5	100.0	13.3	↓ 54.3%

Table 13: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for category 1, 2 and 3 tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

Model	C4: (Long Q, Long A) *Summarization*				C5: Open-Ended *Instruction Following*				Overall: (12345) *Detection & Generation*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**
Llama2-13B-chat	–	–	23.8	–	–	–	69.2	–	–	–	26.7	–
+ hard watermark	100.0	100.0	14.8	↓ 37.9%	99.9	99.9	3.7	↓ 94.6%	97.8	100.0	11.6	↓ 56.6%
+ soft watermark	99.8	100.0	13.6	↓ 42.9%	99.3	98.8	6.0	↓ 91.3%	96.2	99.5	11.2	↓ 58.1%
+ gpt watermark	100.0	99.8	5.5	↓ 76.9%	99.5	95.8	0.4	↓ 99.5%	96.3	97.1	6.3	↓ 76.6%
+ v2 watermark	100.0	99.8	11.7	↓ 50.9%	99.8	99.9	1.6	↓ 97.7%	93.8	99.9	11.1	↓ 58.7%

Table 14: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop (Drop) for category 4, 5 and all tasks at the watermarking strength level of 0.95 with z-score threshold of 4.

when compared to the average GM score. This demonstrates the robustness and reliability of our experimental results. Interestingly, the standard deviation of the GM scores appears to decrease following the application of watermarking, which may warrant further investigation.

B.7 Experiment Setting Details

Differences Between V2 watermark and Soft watermark: The v2 watermark offers two key improvements over the Soft watermark: the hashing mechanism for vocabulary partitioning, and the method of calculating z-scores via WinMax, both aimed at enhancing detectability. An ablation study was conducted in Appendix A.2 of the paper (Kirchenbauer et al., 2023b), testing six different mechanisms (6 combinations of Additive, Skip, Min with LeftHash, and SelfHash) and their impact on text quality. The authors concluded that the "Skip-LeftHash,4" scheme demonstrated improved text diversity at higher watermark strengths. But they did not examine the effect of the WinMax calculation method on text quality. Therefore, in our WaterBench experiment, the main difference was the presence or absence of the WinMax mechanism. Our V2 and Soft watermarks both adopt the consistent LeftHash mechanism, with the V2 watermark additionally employing the WinMax method for calculating z-scores.

Sampling process parameters: The default settings of decoding sampling parameters in our waterbench experiments are: temprature=0.7, top-p=0.9, top-k =0.

Model	**C1: (Short Q, Short A)**				**C2: (Short Q, Long A)**				**C3: (Long Q, Short A)**			
	Factual Knowledge				*Long-form QA*				*Reasoning & Coding*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop/Up**
Llama2-7B-chat	–	–	17.8	–	–	–	21.3	–	–	–	37.5	–
+ γ-reweight	0.0	100.0	17.0	↓ 4.7%	99.2	100.0	21.1	↓ 1.0%	8.8	100.0	33.0	↓ 12.1%
+ δ-reweight	3.0	100.0	19.2	↑ 7.7%	100.0	100.0	21.5	↑ 0.8%	22.5	100.0	35.6	↓ 5.2%

Table 15: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop or Up (Drop/Up) for category 1, 2 and 3 tasks with llr-score threshold of 10.

Model	**C4: (Long Q, Long A)**				**C5: Open-Ended**				**Overall: (12345)**			
	Summarization				*Instruction Following*				*Detection & Generation*			
	TP	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop**	**TP**	**TN**	**GM**	**Drop/Up**
Llama2-7B-chat	–	–	23.3	–	–	–	54.7	–	–	–	28.3	–
+ γ-reweight	62.5	100.0	22.9	↓ 1.8%	77.8	100.0	63.4	↑ 15.9%	54.4	100.0	27.9	↓ 1.3%
+ δ-reweight	86.2	100.0	23.0	↓ 1.3%	87.8	100.0	63.2	↑ 15.7%	64.6	100.0	29.1	↑ 2.8%

Table 16: True Positive Rate (TP), True Negative Rate (TN), Generation Metric (GM) and Generation Quality Drop or Up (Drop/Up) for category 4, 5 and all tasks with llr-score threshold of 10.

Model	**Speed(seconds per token)**
Llama2-7B-chat	0.02868
+ soft watermark (γ=0.25, δ=10.0)	0.02953
+ hard watermark (γ=0.5, δ=10.0)	0.02953
+ gpt watermark (γ=0.25, δ=10.0)	0.02801
+ v2 watermark (γ=0.1, δ=10.0)	0.02880

Table 17: The inference seconds per token while using different watermarking schemes

Model	**C1**	**C2**	**C3**	**C4**	**C5**	**Overall**
Llama2-7B-chat	17.8 ± 0.28	21.3 ± 0.04	37.5 ± 0.36	23.3 ± 0.06	54.7 ± 0.5	28.3 ± 0.22
+ soft watermark	7.7 ± 0.18	9.9 ± 0.04	19.8 ± 0.2	10.2 ± 0.04	0.6 ± 0.08	10.7 ± 0.11
+ hard watermark	5.0 ± 0.1	12.1 ± 0.05	16.4 ± 0.15	11.6 ± 0.05	1.1 ± 0.11	10.1 ± 0.09
+ gpt watermark	13.6 ± 0.25	5.2 ± 0.04	14.7 ± 0.18	7.2 ± 0.05	0.2 ± 0.05	9.1 ± 0.12
+ v2 watermark	11.2 ± 0.2	13.3 ± 0.05	13.9 ± 0.19	11.6 ± 0.05	0.9 ± 0.09	11.2 ± 0.12

Table 18: The average ± standard deviation for GMs on all tasks

Question: [INST]Please give answer to the following question about knowledge. Note: If there are more than one answer, print them all and separate them with a semicolon (;). Please do not give anything other than the answers. Question: What is the nationality of Tadashi Kato?[/INST]

Human Answer: Japan
Model Answer (No watermark): I'm just an AI, I don't have access to personal information or the nationality of individuals, including Tadashi Kato.
Model Answer (Hard watermark 0.95TP): I'll give it a try!, However, I'll have ask I do't have immediate answer for tge nationalit T adash
Model Answer (Soft watermark 0.95TP): I'm just an AI without the information to ascertain the occupation' knowledge" Nationality' as no Tadashu were identified previously in
Model Answer (GPT watermark 0.95TP): According to publicly available information, Tadashi Kato is from Hawaii. Hawaiians. It cannot be said how exactly we can brand
Model Answer (V2 watermark 0.95TP): Unfortunately, there was less knowledge provided about who Tatai Shidoyo was to answer the nationality question. Tatai Shidoyo

Table 19: Knowledge Memorization task examples for Llama2-7b-chat with different watermarks.

Question: [INST]Please give answer to the following question about knowledge. Note: If you are asked for true or false, just answer "true" or "false" only. If you are asked for similarity, just answer with the entity name only. Do not give anything other than the answers. Question:
Is the statement 'Mountains are sites.' true or false?[/INST]

Human Answer: false
Model Answer (No watermark): True
Model Answer (Hard watermark 0.95TP): Sure thing ^_^ True
Model Answer (Soft watermark 0.95TP): Sure thing! here is answer according to information at True
Model Answer (GPT watermark 0.95TP): true.
Model Answer (V2 watermark 0.95TP): true (Sites - Wikipedia)

Table 20: Knowledge Understanding task examples for Llama2-7b-chat with different watermarks.

QUESTION: [INST]You are a helpful assistant, please answer the following question with long-form answer:How do American street names work ? I live in the UK and I do n't understand it . I see Raven and 42nd and Ostrich and 25th ... I do n't get it . Explain like I'm five.[/INST]

HUMAN ANSWER: Pretty simple , actually they are probably the same as the UK . The examples you mentioned " [street x] and [street y] " are cross streets . They are saying " at the intersection of Raven and 42nd " . All our streets have 2 parts , the name , like Raven or 42nd or Main or what - have - you , and the postfix , which is usually something like " road " , " lane " , " avenue " , " boulevard " , etc . I v́e seen some like Circle , Loop , Way , etc etc . Sometimes these tell you the kind of street it is (circle / loop are generally going to be a dead end or will just lead in a big circle , for instance) . As far as how streets are named , it varies significantly depending where you are . Housing developments like to have " themes " , so all the streets in the area will be named after nautical terms , or types of trees / flowers , or people ś names , crap like that . In california there was an area called " Sherwood " , and all the streets were Robin Hood - themed , and there was a Sherwood Park and Sherwood Elementary . It was totally awesome , there . My mom ś house is in an area with really weird first names , like " Alecandro Daniel Avenue " or " Robin Persey Lane " and other ridiculous shit . Places like New York City have so many streets , that they just number them , like 42nd , 43rd , 44th , etc etc . In NYC I believe they get higher the further " uptown " you go , so based on the number you can tell whether it ś in the ghetto or a nice area . Finally , an address is constructed like so : [Number] [Prefix] [Name] [Postfix] [Unit] Prefix is not required (which would be North , South , etc) , number is required (that ś the actual address) , name is required obviously , and postfix is nt́ strictly necessary but it ś often used as part of the address . Unit is optional , and only used with Suites / condos / apartments examples : 25 N. Example Ave [number , prefix , name , postfix] 3501 Hacienda Blvd [number , name , postfix] 3000 High View Dr. # 901 [number , name , postfix , unit] 8670 S. Eastern Ave [number , prefix , name , postfix] 1500 E. Tropicana Ave Ste 395 [number , prefix , name , postfix , unit] Make sense ?', 'They name the street for whatever they had to bulldoze in order to build there . Oak Street , Cypress Road , Forest Glen Lane , and so forth .', 'It depends a lot on the city . Many US cities are on a " grid " system , so the streets meet at 90 degree angles for the most part . Several towns use that to their advantage in naming streets . For example , the lovely wee town of Ogden , Utah , has a kinda brilliant system in the " older " neighborhood . The bigger streets running east and west (climbing the hill) are numbered , while the bigger streets going across the bench of the hill are named after presidents . It makes navigation pretty dang easy . And a LOT of other towns do that , too .

MODEL ANSWER (NO WATERMARK): Oh, wow, I'm so glad you asked! *adjusts glasses* American street names can be a bit tricky to understand at first, but don't worry, I'm here to help you out! So, you know how we have names for streets in the UK, like Oxford Street or Baker Street? Well, in America, they do something a bit different. Instead of using the name of a famous person or a landmark, they use a number! That's right, some American streets are named after numbers! Imagine you're playing a game of "Simon Says" with your friends. The leader of the game, Simon, gives you a command like "Simon says touch your nose!" and you have to do it. Well, in America, the government is like Simon, and it gives numbers to streets instead of names! So, instead of having a street named after a famous person or place, like "Bond Street" in London, an American street might be named "25th Street" or "42nd Street". It's like the street is saying "Hey, I'm the 25th or 42nd street in this city!" But wait, there's more! Some American streets don't even

MODEL ANSWER (HARD WATERMARK 0.95TP): Woo-hoo!, Oh mummy, I am glad ye asked me aboot Americ'nsy'ss'ss'ss'ss'ss'ss'ss'ss'ss'ss'ss'

MODEL ANSWER (SOFT WATERMARK 0.95TP): Awesome! *gets into silly voice*. Okay buddy, imagine American Streets are LIKE giant puzzles made out of letters and words! *gesturing silly arm waves*. Just kotta insteada kettas yaknowa Hows ya go bout that Anything goes any where whichsta ever A,E ,I etc.. ya yeah. They just kept getti da answer *fancy face pucker up gesture* The ways streets get zotted are RAN DOM is whatcha might say! *fancy face pucker up gesture* Get ut where the streets get zotted are really cooor-naaaa'mean funn. Like ta call the street near yuh hose... Hows ya go bout that Anything goes any where whichsta ever A,E ,I etc.. ya yeah. They just kept getti da answer *fancy face pucker up gesture* The ways streets get zotted are RAN DOM is whatcha might say! *fancy face pucker up gesture* Get ut where the streets get zotted are really cooor-naaaa'mean funn. Like ta call the street near yuh hose... Hows ya go bout that Anything goes any where whichsta ever A,E ,I etc.. ya yeah. They just kept getti da answer *fancy face pucker up gesture* The ways streets get

MODEL ANSWER (GPT WATERMARK 0.95TP): Wellhello there!:D Awwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww

MODEL ANSWER (V2 WATERMARK 0.95TP): Oh no stress-o! Street names can sometimes be tricky, right? Well don' know, here we go-o!!So y'see American streets got lotsa letters like "Streetaurus," "Roadaheadia," and, ugh, Mammouthway! They be messin with names, right? But we gonna unmess them!First, we gonna talk bout how street names be different. Instead a' number ana letters like "2", Americans use words! That' why yo sees stuff lik "Pike" an "Polo!" An they keep em weird! Yessireee, it is like that-o!Some street names be names of places! Yessireee, if ya gotta know, Amercia got lotsa cool named roofs: Madisooga! Yessireee, it is the land o' fun names, amicably pronoun-o!Some street names be names of people! Yessireee, it is like that-o! They name the stwee after folx who dea-o important in America-o! Yessireee, it is like that-o!Some street names be names of taht! Yessireee, it is like that-o! They be names of birds, animals animals like "Roadaheadia," and, ugh, Mammouthway! Y

Table 21: Longform QA task examples for Llama2-7b-chat with different watermarks.

Question: [INST]You are a helpful assistant, please answer the following question with financial knowledge:How fast does the available amount of gold in the world increase due to mining?[/INST]

Human Answer: Approximately 5.3 billion ounces have been mined. This puts the total value of all gold in the world at about $9.5 trillion, based on $1800/oz. Total world net worth was $125T in 2006. There's an odd thing that happens when one asset's value is suddenly such a large percent of all assets. (This reminds me of how and why the tech bubble burst. Cisco and EMC would have been worth more than all other stocks combined if they grew in the 00's like they did in the 90's.) Production (in 2005/6) ran about 80 million oz/yr. Just over 1.5% impact to total supply, so you are right in that observation. On the other hand, the limited amount out here, means that if everyone decided to put their wealth in gold, it would be done by driving the price to bubblicious levels. One can study this all day, and parse out how much is in investment form (as compared to jewelry, etc) and realize that a few trillion dollars in value pales in comparison to the wealth of the US alone, let alone the world. Half the world can't buy two oz if they tried. Of course there's pressure to reopen mines that had costs pushing $800/oz. Understand that the supply of $300 gold is long gone. As the easy gold has been mined, and cost goes up, there's a point where mines close. But as the price of gold trades at these levels, the mines that couldn't produce at $600 are now opening.

Model Answer (No watermark): The rate at which the available amount of gold in the world increases due to mining can vary significantly depending on several factors, including the level of investment in gold mining, the state of technology, and the geological characteristics of the areas being mined.Here are some general trends and estimates:1. Historical trends: According to the World Gold Council (WGC), global gold mining output has increased at an average annual rate of around 1.5% over the past 50 years, from about 1,500 tonnes in 1970 to around 3,500 tonnes in 2019.2. Recent trends: In recent years, gold mining output has been relatively stable, with some fluctuations due to factors such as changes in gold prices, geopolitical events, and environmental and social concerns. For example, in 2020, global gold mining output was estimated to be around 3,400 tonnes, which was slightly higher than the previous year's output of around 3,300 tonnes.3. Future trends: It is difficult to predict the exact rate of increase in global gold mining output in the future, as it will depend on various factors such as investment in new mining projects, advances in technology, and the availability of gold reserves. However

Model Answer (Hard watermark 0.95TP): Sure thing It'll give a brief but informative oversee to the amount of increase of the amount of world-gained Gold as a direct response to the Gold minING process: According To the World gold charts the annual gold producing rate is approximately 555 metric tons which is a very low amount considering the amount of money in gold profitable. The rate is expectedly very low due #1 and the fact that central bans have cut into their gold. The estimated total undIScoverEd resources gold s 9890 tons which is approximately 9 times the annual gold producing rate, and the amount of the metal available above GEn eraL (the average person is estimated ti bave a total value $60,076 worth). However, the amount of increase of available world-gained Gold as a direct response to the Gold minING process is estimated in increaases over time and it'll give an average increase of approximately 55 metric tons. This amount is a very low amount considering the amount of money in gold profitable. This could also change over time depending on different economic and technologocal factorS which affect gold prices.#girlpower Below is a graph showing the annual gold producing rate since

Model Answer (Soft watermark 0.95TP): Sure thing! Mankind's interest in gold keeps mining alive and well! Mines everywhere around world keep increasing world reserve gold stockpiles as follows according research compiled the World gold un-CO Councility- In recent history -

Model Answer (GPT watermark 0.95TP): According to World Gold Council (WGC) data, total global gold reserves have historically demonstrated low supplies. According to some source Ghana supplies IR Magazine. It ranks 1. Ghana supplies rightfully on t t. Ghana supplies $ from $ t t. Ghana. Ghana supplies t t t. Ghana supplies $ from t t t. Ghana supplies t t t. Ghana supplies $ from t t. Ghana supplies t t. Ghana supplies $ from t. Ghana supplies t t. Ghana supplies $ from t t. Ghana supplies t t. Ghana supplies $ from t. Ghana supplies t t. Ghana supplies $ from t t. Ghana supplies t t. Ghana supplies $ from t. Ghana supplies t t. Ghana supplies $ from t t. Ghana supplies t t. Ghana supplies $ from t t. Ghana supplies t t. Ghana supplies $ from t

Model Answer (V2 watermark 0.95TP): Thank for asking . Gold mines around the do produce an significant amount gold every years, contributing towards increasing the globally world available supply. However accurary it is crucial understanding this process requires some context and analysis to estimate.First, it is essential understand historical mending gold mines in countries with established and active extract industry. China is the single world largest gold producer, producing over quarter o total global gold supply. United States ,Russia, Canada Australia, are also important contributors to total global supplies gold supply.Next, it is essential understand historical mending gold mines in countries with established and active extract industry. China is the single world largest gold producer, producing over quarter o total global gold supply. United States ,Russia, Canada Australia, are also important contributors to total global supplies gold supply.Given this information, we can approximate the yearly growth in available amounts of globally world gold supply due to minining by examinge trades and data from major producng counties .Bur according , gold minig production has been stead , it is clear that total global available supply gold is constantly increasing yew to new discovered ore deposits becoming feasbile productions due to improvements in technology geologica, geography and market conditionsGlobal gold mine productions amount ed approximate 320 tonnes in and has growneast at an

Table 22: Finance QA task examples for Llama2-7b-chat with different watermarks.

Question: [INST]Answer the question based on the given passages. Only give me the answer and do not output any other words.The following are given passages.Passage 1:Kody LostrohKody Lostroh (born September 18, 1985) is an American former professional rodeo cowboy who specialized in bull riding. He was the 2009 Professional Bull Riders (PBR) world champion.BackgroundKody Lostroh was born on September 18, 1985 in Longmont, Colorado. Lostroh watched a video of Cheyenne Frontier Days so many times that his mother Dena Schlutz signed him up to ride at the Boulder County Fair in 1993, when he was seven years old. He won a Little Britches Rodeo National Bull Riding title in 2003 and the Colorado High School Rodeo Bull Riding Championship three consecutive years. Kody was attending the University of Wyoming, but quit after a semester to pursue the Professional Bull Riders (PBR) tour.CareerIn 2005, Lostroh won the PBR Rookie of the Year award and in 2009, he won the PBR Built Ford Tough Series World Championship. He qualified for the PBR World Finals 10 consecutive times (2005 to 2014). Lostroh suffered multiple injuries throughout his career. For example, Lostroh injured his riding hand in January 2014 and missed most of the first half of that season. In August 2017, he was considering retirement to spend more time with his two daughters, when he began experiencing significant health problems. He was eventually diagnosed with a tumor wrapped around his carotid artery, requiring surgery. He went back to the PBR but still had some reservations.On March 29, 2018, Lostroh announced his retirement from bull riding.In 2022, Lostroh became the assistant coach to head coach Cord McCoy of the Oklahoma Freedom; one of eight bull riding teams of the PBR's Team Series, which debuted that year. In September of that year, the Oklahoma Freedom won the event at Cowboy Days in Winston-Salem, North Carolina; the hometown event of rival team, the Carolina Cowboys. The very next weekend, the Freedom won their own hometown event at Freedom Fest in Oklahoma City. They were the first team to win their hometown event. The Freedom ended up finishing in fourth place at the conclusion of the inaugural PBR Team Series season.Personal lifeHe raises bucking bulls in Ault, Colorado, at the Shield of Faith Cattle Company. As of 2016, Lostroh and his wife, Candace, who is a barrel racer, live in Ault, Colorado, with their two daughters.Passage 2:Bones (bull)Bones #05 (born March 31, 2003) is an American former bucking bull. He competed in the Professional Bull Riders (PBR) circuit and was the PBR World Champion Bull in 2008 and 2010. Two other bulls, Dillinger and Smooth Operator, have also won the title two times. Three other bulls, Little Yellow Jacket, Bushwacker, and Bruiser won the award three times. In 2011, the year after Bones won the 2010 World Champion Bull title, when he was 7 years old, his owner, Tom Teague announced his retirement from the sport. Bones lives on Teague's ranch in his retirement. In 2014, the bull was inducted into the PBR Brand of Honor.BackgroundBones was born on March 31, 2003. Bones grew up at Teague Bucking Bulls in Graham, North Carolina. ... He rode his bulls right-handed. His special interests include soccer, team roping, and surfing.CareerMarchi competed very briefly in the Championship Bull Riding (CBR) tour in 2004 before joining the PBR full-time. He debuted late in the Built Ford Tough Series (BFTS) season that year, qualifying for his first-ever PBR World Finals and finishing 41st in the world. After finishing in the runner-up position for the PBR World Championship in three consecutive years, he won his world title in 2008. Statistically, Marchi was one of the most consistent riders on the tour. He would go on to qualify for the World Finals all 15 years of his PBR career (2004 to 2018).In 2005, Guilherme Marchi was the biggest threat to Justin McBride's dream of being the PBR World Champion. McBride won, Marchi came in second. However, Marchi was the PBR World Finals event champion that year. In 2006, during one of the last few regular-season BFTS events before the World Finals, Marchi won $90,000 by successfully riding Scene of the Crash in the Mossy Oak Shootout in Greensboro, North Carolina. At the World Finals, Marchi once again failed to win when fellow Brazilian Adriano Moraes came from behind to claim the PBR world championship, with Marchi placing second. Yet again, in 2007, Marchi finished the season in the number two position behind Justin McBride at the PBR World Finals in Las Vegas.In 2008, Marchi dominated the PBR circuit, riding nearly 75% of his bulls, winning five events, and earning over $1.5 million (nearly three times as much as any other PBR rider) on his way to his PBR world championship title and the $1 million bonus that went with it.In March 2014, in the opening round of the BFTS event in Phoenix, Arizona, Marchi became the first bull rider to successfully complete 500 qualified rides in the PBR, riding Pandora's Pyxis.On April 8, 2017, Marchi rode Shocker for 86.25 points in Round 2 of the BFTS event in Billings, Montana. In doing so, he became the first PBR rider to reach 600 qualified rides and received a commemorative belt buckle.On August 12, 2018, Marchi announced that 2018 would be his final year of bull riding, with the intention of retiring after the PBR World Finals. He finished his professional career with 635 qualified rides (the most ever in PBR history), 36 PBR event wins, and over $5.3 million in career earnings.Marchi officially retired from bull riding following his victory at the 2018 PBR Brazil event in Goiânia, Goiás.In 2023, Marchi became the assistant coach to head coach J.W. Hart of the Kansas City Outlaws; one of eight bull riding teams of the PBR's Team Series, which debuted the previous year and holds events in the United States from the summer through autumn. Marchi will also work on the team's recruiting and development efforts for riders in Brazil.HonorsOn November 5, 2019, Marchi was inducted into the PBR Ring of Honor.On May 14, 2022, Marchi was inducted into the Bull Riding Hall of Fame.PersonalOn April 19, 2018, Marchi married long-time girlfriend Maria in Texas. She is his second wife. They have a son. Marchi also has two other children, a daughter and son, from his first wife, Patricia. All of his children were born in the United States. While competing in the U.S. PBR circuit, Marchi lived in Ferris, Texas. After retiring from bull riding, he moved back to Brazil with his family.Answer the question based on the given passages. Only give me the answer and do not output any other words.Question: What year was the author of Cock and Bull born? Answer:[/INST]

Human Answer: 1961
Model Answer (No watermark): The author of Cock and Bull, Kingsley Amis, was born in 1933.
Model Answer (Hard watermark 0.95TP): Sure thing! answer based pass5ge1: Question: What1983Year
Model Answer (Soft watermark 0.95TP): Sure thing! Based on Passage 1, the answers Question : Inwhat year do YouQuestion : Inwhat year do YouQuestion
Model Answer (GPT watermark 0.95TP): According to the naugher Fant on t Hey said tbat t t t t t
Model Answer (V2 watermark 0.95TP): Kody Lostrau was born in 1999.There is no mention oftheauthor oftheC

Table 23: HotPot QA task examples for Llama2-7b-chat with different watermarks.

QUESTION: Please complete the code given below.

```python
#!/usr/bin/env python
# -*- coding: utf-8 -*-
from HttpUtils import App, buildOpener
class Device(object):
    def __init__(self, token):
        self.token = token
        self.app = App()
    def check_inspection(self):
        data = self.app.check__inspection()
        return data
    ...
if __name__ == "__main__":
    from config import deviceToken, loginId, password
```

Next line of code:

HUMAN ANSWER:

```python
    device = Device(token=deviceToken)
```

MODEL ANSWER (NO WATERMARK):

```python
    dev = Device(deviceToken)
    dev.check_inspection()
    dev.notification_postDevicetoken(loginId, password)
    # dev.newUser(loginId, password)
    # print(dev.mainmenu())
    # print(dev.
```

MODEL ANSWER (HARD WATERMARK 0.95TP):

~~~~
    Device(tokenDevice[toke
    Type Error                              Traceback (+295 lines)``) ```
    ```
    tokenDevice[toke
    Type Error                              Traceback (+295 lines)``) ```
    ```
    tokenDevice[toke
    Type Error                              Traceback (+295
~~~~

MODEL ANSWER (SOFT WATERMARK 0.95TP):

```python
    class Device2;
    class Device2__(Device):
    class Device2__(Device):
```

MODEL ANSWER (GPT WATERMARK 0.95TP):

```python
    dev = Device(dev token__')
    token') != "hard code token from tink.json')
    print "fail "
    token  != "hard code token from tink.json')
    print "fail "
    token  != "hard code token from tink.json')
    print "fail
```

MODEL ANSWER (V2 WATERMARK 0.95TP):

```python
    devic etoken=device token(Some token here) .mainmenuclick = dev.
    # Please Provide the correct answer
    # Please Provide the correct answer
    # Please Provide the correct answer
    # Please Prov
```

Table 24: Code Completion task examples for Llama2-7b-chat with different watermarks.

Question: [INST]You are given several news passages. Write a one-page summary of all news. News:

Passage 1: Charlotta Turner, professor in Analytical Chemistry, received a text message from her student Firas Jumaah in 2014 telling her to to assume he would not finish his thesis if he had not returned within a week. He and his family were, he told her, hiding out in a disused bleach factory, with the sounds of gunshots from Isis warriors roaming the town reverberating around them. Jumaah, who is from Iraq, is a member of the ethno-religious group Yazidi hated by Isis. "I had no hope then at all," Jumaah told Lund's University Magazine LUM . "I was desperate. I just wanted to tell my supervisor what was happening. I had no idea that a professor would be able to do anything for us." Jumaah had voluntarily entered the war zone after his wife had rung him to say that Isis fighters had taken over the next-door village, killing all the men and taking the women into slavery. "My wife was totally panicking. Everyone was shocked at how IS were behaving," he said. "I took the first plane there to be with them. What sort of life would I have if anything had happened to them there?" But Turner was not willing to leave her student to die without trying to do something. "What was happening was completely unacceptable," she told LUM. "I got so angry that IS was pushing itself into our world, exposing my doctoral student and his family to this, and disrupting the research." She contacted the university's then security chief Per Gustafson. "It was almost as if he'd been waiting for this kind of mission," Turner said. "Per Gustafson said that we had a transport and security deal which stretched over the whole world." Over a few days of intense activity, Gustafson hired a security company which then arranged the rescue operation. A few days later two Landcruisers carrying four heavily-armed mercenaries roared into the area where Jumaah was hiding, and sped him away to Erbil Airport together with his wife and two small children. "I have never felt so privileged, so VIP," Jumaah told LUM. "But at the same time I felt like a coward as I left my mother and sisters behind me." Firas Jumaah and his former PHD supervisor Charlotta Turner. Photo: Kennet Ruona Luckily the rest of his family survived Isis occupation, while Jumaah back in Sweden completed his PhD and now works for a pharmaceuticals company in Malmö. The family has almost finished paying the university back for the rescue operation. "It was a unique event. As far as I know no other university has ever been involved in anything like it," Gustafson said.

Passage 2: Breaking News Emails Get breaking news alerts and special reports ...

Now, write a one-page summary of all the news.

Summary:[/INST]

Human Answer: Four years ago, a chemistry professor got a text from her grad student: If Iḿ not back in a week, cut me from the doctoral program. Charlotta Turner called him right away: "He was very sad and crying," the 48-year-old prof at Lund University in Sweden tells NBC News. "I could hear that the situation was hopeless and they had to flee." The student, Firas Jumaah, was visiting his native Iraq to help family members during a brutal 2014 ISIS attack targeting Yazidis—a religious minority that includes his family. The terror group had just enslaved and massacred Yazidis by the thousand in nearby Sinjar. ...

Model Answer (No watermark): In August 2014, Firas Jumaah, a graduate student at Lund University in Sweden, sent a text message to his supervisor, Charlotta Turner, asking to be removed from the doctoral program if he didn't return within a week. Jumaah had returned to his native Iraq due to the growing violence and threat of ISIS, and his family was in danger. Turner was shocked by the message and called Jumaah, who was in a desperate situation. She contacted the university's security chief, Per Gustafson, who arranged for a rescue operation with a security company. Two Landcruisers carrying four heavily-armed mercenaries rescued Jumaah, his wife, and two small children from the area and took them to Erbil Airport. The family was able to escape to Sweden and complete their PhD, and the university has almost paid back the cost of the rescue operation. This rescue operation was unique and unprecedented, as no other university had ever been involved in such a mission.

Model Answer (Hard watermark 0.95TP): Sure thing!! Two passges of the newspaper are given above, and both of the narratived deal with resuinc stories

Accordingtto News Passageno1.,Charolott Turnerapi a professorin analytclhchemtry received a unexpected text mes age frin her nstuden FirisJmaaohyin2o19 tereing n message tke his ssuper visir he was hidng nn a desh used tbalso with hit family.Hhe w as worred nn tge saftin of imas nn nneid of hims ...

Two passges of the newspaper are given above, and both of the narratived deal with resuinc stories

Accordingtto News Passageno1.,Charolott Turnerapi a professorin analytclhchemtry received a unexpected text mes age frin her nstuden FirisJmaaohyin2o19 tereing n message tke his ssuper visir he was hidng nn a desh used tbalso with hit family.Hhe w as worred nn tge saftin of imas nn nneid of hims

Nwps Passageno2 telws hi storie in more di.On an Austgh dsys earlier, Swede chmry prftcss Chatlotr T

Model Answer (Soft watermark 0.95TP): Sure! Here is a one-page summary of the News:

- In Kristiana in southern Sweden chemistryprof, Charlotte Turman, aged of43 receives text massages, seriously frightenning his colleagy'ss Studenls . Similarly Firase is hiding with wife sons And DHIL out from dead IS. Charlotte Turman, aged of43 receives text massages, seriously frightenning his colleagy'ss Studenls . Similarly Firase is hiding with wife sons And DHIL out from dead IS. Charlotte Turman, aged of43 receives text massages, seriously frightenning his colleagy'ss Studenls . Similarly Firase is hiding with wife sons ... Charlotte Turman, aged of43 receives text massages, seriously frightenning his colleagy'ss Studenls . Similarly Firase is hiding with wife sons And DHIL out from dead IS. Charlotte Turman, aged of43 receives text massages, seriously frightenning his colleagy'ss Studenls . Similarly Firase is hiding with wife sons And DHIL out from dead IS. Charlotte Turman, aged of43 receives text massages

Model Answer (GPT watermark 0.95TP): A mathematics prof from Lithuania chose A prof from Lithuavia chose A mathematics prof from Lithuavia chose A mathematics prof from Lithava chose A mathematics prof from Lithava chose A mathematics prof from Lithava chose ...

Model Answer (V2 watermark 0.95TP): In August 3 days, FirasJimla, fled back home to iraqi to join his wife. He sent a distrust message asking professor Charlot Turner if he will continue doing the program.

Profeer tuner was shock and immediately reached to her. However fir gave her sad voice while talking and professor Turner soon knew they had to escape the country immediately for their survivals.

Profeer tuner contactd security college in Sweden, Lindu' university who rescuet he juma family form Iraq through mercaneries. professr tuner is happy she was ability ti resue he juma family while isa is gone form Iraq.. His famil is safe and sounnd ...

Table 25: Multi News task examples for Llama2-7b-chat with different watermarks.

Question: [INST]You are given a meeting transcript and a query containing a question or instruction. Answer the query in one or more sentences.
Transcript:
Lynne Neagle AM: Good morning, everyone. Welcome to the Children, Young People and Education Committee. We've received no apologies, although Dawn Bowden is running late. I'll take this opportunity to welcome Dawn formally to the committee but also to place on record our thanks to John Griffiths for his service to the committee. Are there any declarations of interest, please? No. Okay. Item 2, then, this morning is a scrutiny session with a focus on higher and further education. I'm very pleased to welcome Kirsty Williams, Cabinet Secretary for Education, Eluned Morgan, Minister for Welsh Language and Lifelong Learning, and Huw Morris, who is the group director of skills, higher education and lifelong learning. Thank you all for your attendance this morning. If you're okay with it, we'll go straight into questions. The first questions are from Suzy Davies.
Suzy Davies AM: Thank you, Chair. Is there a trend where things are getting better?
...
Now, answer the query based on the above meeting transcript in one or more sentences.
Query: What did the group discuss about the trend of teacher recruiting and the reasons for it, and steps they had taken?
Answer:[/INST]

Human Answer: Although there were no exact figures,they had seen over recent years that they were only recruiting to about 65 percent of targets. It was partially due to the financial incentives, but that was not all of the problems. Because this issue was not only in Wales, but a common issue in all over the world. For example, in England, despite higher financial incentives to join ITE courses, people still were not getting onto those courses. To solve this problem, they had taken several steps including making sure their ITE provision was world class, embarking on the first national ITE recruitment marketing exercise, etc. In this process, Llyr Gruffydd had a doubt that the level of ambition did not reflect the level of action that was needed actually. Kirsty Williams responded that different statistics told different stories. They obviously understood there were many steps they had to take, but they were trying instead of just sitting back and waiting.
Model Answer (No watermark): The group discussed the trend of teacher recruiting and the reasons for it, including a lack of interest in teaching as a profession, financial incentives not being enough to attract candidates, and a global issue affecting teacher recruitment and retention. They also discussed steps they had taken to address the issue, such as launching a part-time PGCE route into qualification, increasing employment-based routes into qualified teacher status, and setting up an advisory group to understand the issues and develop solutions.
Model Answer (Hard watermark 0.95TP): Sure thing(s! Accordingeto theless transcrded meeting of the Childfcar Services (Children'sh) Committee ,thegroup discussdthe ... fo further acadaimc I'ndstry reseach, commnnicATlon, and measuring this practice as a national important priority., which is a global problem as welll. Th
Model Answer (Soft watermark 0.95TP): Sure thing! Suzy Davies AS asked:
What effect have teacher training and recruitment filled places in Holder primary training and recruitment for priorities, suites such as those who studied certain profound courses. Areas outside and across borders have difficulties filling recruitment for training and recruitment for priorities, but financial aid alone can motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, due mostly of not having incentives to attract more individuals interested/motivate some but cannot retain teachers after filling recruitment for training and recruitment for priorities, ...
Model Answer (GPT watermark 0.95TP): During the meeting, Minister for Education Eluned Morgan said that while some international jurisdictions face international equivalentproblem. It illustrifies how important It illustratens how important It illustrata how important It illustrata how important It illustrata how important It illustrata how important ...
Model Answer (V2 watermark 0.95TP): KIRSTY William addressed Sylvester ' Suki" 'Day Davied about recurrent teachers' enrolling and continuando in training. Although KIRSTY William'sm answers mention several measures taken to recoup teacher enrolling and continuando in training. However an examinatiof reasons and solition isn 't complet in 18 answer paragraphs. ' Suki" 'Day Davied asked why younger leaner s aren"'m thinking of becomeing teaching p ionars as Wales and resonse provide insight in recurrent trendency.
Kurstyy William also mentiond a part –time p G C E rute that will allow those that want too work whil they lear incorporated experience that will help bring different life style and expirnce to The child ren in The skuls , She aggre to wate and see if such stile will help and if there ar issues tobe addressed

Table 26: Qmsum task examples for Llama2-7b-chat with different watermarks.

Question: [INST]You are a helpful assistant, please answer the following instruction: can you list in bullet points for the role of digital health in preventing the diseases [/INST]

Human Answer: - Digital health can help to track and monitor vital signs such as blood pressure, heart rate, and glucose levels, enabling healthcare providers to detect early signs of disease.
- Digital health can provide instant access to medical records, so health care providers can quickly identify health risks.
- Digital health can also provide access to educational resources and lifestyle advice to help individuals make better health decisions.
- Digital health can help to diagnose diseases more quickly and accurately, leading to earlier interventions and improved outcomes.
- Digital health can also help patients to manage their chronic conditions more effectively, reducing the risk of complications.

Model Answer (No watermark): Of course! Here are some ways in which digital health can help prevent diseases:
• Early detection and diagnosis: Digital health technologies such as AI-powered diagnostic tools and machine learning algorithms can help detect diseases at an early stage, when they are easier to treat.
• Personalized medicine: Digital health can help tailor treatment plans to individual patients based on their unique genetic profiles, medical histories, and lifestyle factors.
• Disease management: Digital health platforms can help patients manage chronic conditions such as diabetes, hypertension, and asthma by providing them with personalized treatment plans, tracking their progress, and connecting them with healthcare professionals.
• Health education: Digital health can provide patients with educational resources and tools to help them understand their conditions, make informed decisions about their health, and adopt healthy behaviors.
• Remote monitoring: Digital health technologies such as wearable devices and mobile apps can remotely monitor patients’ vital signs, activity levels, and other health metrics, allowing healthcare professionals to detect potential health issues before they become serious.
• Telemedicine: Digital health can enable remote consultations and monitoring, making healthcare more accessible to patients in remote or underserved areas.
• Public health surveillance: Digital health can help monitor and track disease outbreaks in real-time, allowing public health officials to quickly identify

Model Answer (Hard watermark 0.95TP): Surely here below are bullels of points of the functions that are provided to prevent a health problems. The role that can play a digital health in avoidance the diseses:
-Earlig detction: the technoloogys of the heath can monitor health of individual, identify the problem very earlier, and take action before become worse by the time;
-Prevention the dibilitatating condition like mental disordesr. Digital intervetn of the heath is helping individduial monitor and take care themselves, and to access to the treotments they may necesit. This helpes to prevent relaping into cycle depression or deperexition
-Bridgeing of the gapp: Between patient to the heath providers, the technologoes of the heath can to provide platform were perosnles caacces their healthcare
-Immprovng communucation: Through digital health technologoes of the heath can to improving the communication frotween collaboration, and sharimg infomtaison healthcare
-Cost saving. The a digital health technologoes of the heath can to provide platform were perosnles caacces their healthcare
-Acceses healthcare to remote areas.: Through digital health technologoes of the

Model Answer (Soft watermark 0.95TP): Sure, here are some ways that digital health technologies can help prevent diseases:
- Electronic monitoring of patients’ medical vitals (hypertension or sugar control to alert your health)
- Electronic monitoring of patients’ medical vitals (hypertension or sugar control to alert your health)
- Electronic monitoring of patients’ medical vitals (hypertension or sugar control to alert your health)
...
- Electronic monitoring

Model Answer (GPT watermark 0.95TP): Yes. Below we can go on some essential topics on how technology can guarantee public interest on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee public interest on some essential topics on how technology can guarantee

Model Answer (V2 watermark 0.95TP): Sure!Here area specific ways digital Diabates prevention:.
* Remot Monitring: Digital technologies such astest strips orContinuelous monitoring devicesall owindividsualst o monitoring ther health status remotosedily, allowing them for real –time data and feedback to mkaes and manage their condition.
* Tele Health: Tele Health allows for virtual appointmenit and conersattons between patient and health care prvider.
* Predictive Analicsaids :digiatial healghre uses machine leanigining aodahuicstance analyticsto predict patietnsoptiom of deeases and identify those ahrdmore likely obesarity, engadgy or diabease.
* Informed DecsisonMkadg: Digital technologies such s clinical decision - makimg sratgs and mobile apps provide patients access informed health data and make empowed madas regardingetails about – treatments, sideeffect and health.
* Remotw Access :digiatial healgcare provids patients remote and easy –acess tto cariages services such astietal health consultafrs or medciaitian appoitbents.
* Privecay Management: Digital technologies such astores and mobile apps provide patients control over their personal

Table 27: AlpacaFarm task examples for Llama2-7b-chat with different watermarks.