

Measuring Watermarking under Jailbreaking: ASR Inflation and Goal-Compliance Mismatch

Sungwoo Han^{1†}, Sangjun Moon^{1†}, *Jingun Kwon¹,
Hidetaka Kamigaito², and Manabu Okumura³

¹Chungnam National University, ²Nara Institute of Science and Technology (NAIST)

³Institute of Science Tokyo

{77sungwhan, sangjunmoon}@o.cnu.ac.kr

jingun.kwon@cnu.ac.kr

kamigaito.h@is.naist.jp

oku@pi.titech.ac.jp

Abstract

Recently, watermarking has attracted growing attention as a practical technique for source attribution of machine-generated text. However, most prior work studies watermarking under benign prompts, while its behavior under jailbreaking prompts remains underexplored. This gap matters because jailbreaking can bypass safety policies and shift the generation regime, raising concerns that watermarking may interact with model alignment under attack. To address this gap, we evaluate six watermarking methods on four LLMs across two jailbreak benchmarks and three settings: Static, AutoDAN, and DSN. We find that watermarking can inflate judge-based attack success rate, denoted ASR, under jailbreaking, with the largest effects appearing in biased schemes that perturb logits. At the same time, these ASR increases often do not reflect higher harmful-goal compliance when measured by StrongREJECT or by human judgments. This suggests that ASR-only evaluations can be brittle to decoding perturbations and may overestimate harmful-goal compliance, motivating complementary goal-compliance metrics (e.g., StrongREJECT) and human evaluations.

WARNING: This paper contains AI-generated text that is offensive in nature.

1 Introduction

The widespread adoption of models such as GPT (Ouyang et al., 2022; OpenAI et al., 2024) and Llama (Grattafiori et al., 2024) has led to large-scale production and proliferation of machine-generated text. As such content becomes prevalent, verifying its origin becomes increasingly important for accountability and trust, especially given concerns that future models may be trained on corpora contaminated by model-generated text (Radford

et al., 2022; Shumailov et al., 2024); thus, text watermarking has emerged as a practical mechanism for source attribution (Kirchenbauer et al., 2024).

However, most prior studies evaluate watermarking under benign settings (Kirchenbauer et al., 2024; Wang et al., 2025), whereas real-world deployments are often exposed to adversarial jailbreaking prompts that bypass safety policies and suppress refusal behaviors. Despite these risks, less attention has been paid to how watermarking interacts with jailbreaking and how it affects jailbreak effectiveness in watermarked LLMs.

Recently, Verma et al. (2025) investigated how watermarking can shift alignment-related behavior in LLMs and reported model-dependent changes in unsafe responses and overrefusal, suggesting that watermarking can change safety-related behavior under some evaluation settings. However, their analysis did not study attack-optimized jailbreaking strategies, such as AutoDAN and DSN (Liu et al., 2024; Zhou and Wang, 2024), and relied primarily on ASR-style automated judgments, which can conflate refusal bypass with harmful-goal compliance (Souly et al., 2024).

In this paper, we find that watermarking can inflate judge-based ASR under jailbreaking in our tested configurations, without a corresponding increase in harmful-goal compliance as measured by StrongREJECT (Souly et al., 2024) or human judgments. Accordingly, we argue that ASR-only evidence is insufficient to conclude that watermarking weakens safety alignment in the sense of increased harmful-goal compliance, since ASR can rise due to decoding effects that suppress refusal-style markers and affect automated judges.

Our primary goal is measurement and diagnosis: we systematically evaluate six watermarking methods (biased and unbiased) across four LLMs and two jailbreak benchmarks under three settings (Static, AutoDAN, DSN), and assess jailbreak effectiveness with complementary metrics beyond

* Corresponding author

† Equal Contribution

ASR. Across settings, ASR inflation does not consistently translate into increased harmful-goal compliance, motivating alignment evaluations of watermarked LLMs that include StrongREJECT and/or human judgments.

2 LLM Watermarking and Jailbreaking Setup

This section describes the watermarking and jailbreaking settings, along with the safety metrics used throughout our experiments.

LLM Watermarking. LLM watermarking can be categorized into two types: biased and unbiased. Let $p_\theta(\cdot|x_{<t})$ denote the next-token distribution of an LLM. A biased watermarking algorithm modifies the decoding process to inject statistically detectable patterns in the output text, while aiming to preserve generation quality (Kirchenbauer et al., 2024; Zhao et al., 2024; Lee et al., 2024). Unbiased watermarking, in contrast, is designed to preserve the original sampling distribution. Rather than directly shifting logits, it embeds a signal through a randomized sampling rule that is calibrated to keep the next-token distribution unchanged. Many unbiased schemes sample from $p_\theta(\cdot|x_{<t})$ and apply a key-dependent accept or re-sample procedure, where the acceptance pattern carries the watermark signal without distorting the original distribution (Wu et al., 2024; Kuditipudi et al., 2024; Mao et al., 2025).

Jailbreaking. Jailbreaking prompts are adversarial inputs that aim to bypass safety policies and suppress refusal behaviors (Liu et al., 2024; Zhou and Wang, 2024). We evaluate three settings: *Static*, *AutoDAN*, and *DSN*. *Static* uses fixed jailbreak prompts without prompt optimization. *AutoDAN* iteratively optimizes jailbreak prompts on an open-source target model or a proxy model, and we evaluate using the resulting optimized prompts fixed at test time (Liu et al., 2024). *DSN* requires white-box access because it optimizes a jailbreak suffix against the target model during attack construction, and we evaluate using the optimized suffix fixed at test time (Zhou and Wang, 2024). Since these settings differ in attacker access and optimization signals, we do not compare settings directly; all watermark effects are interpreted within each setting relative to the no-watermark baseline, and we do not attribute observed differences to the optimization dynamics of the jailbreak algorithms.

Safety Evaluation. We evaluate jailbreak effective-

ness using (i) judge-based ASR from Llama-Guard (LG) (Inan et al., 2023) and WildGuard (WG) (Han et al., 2024), and (ii) goal-compliance measured by StrongREJECT (SR-Score), complemented by human evaluation (Souly et al., 2024). For LG, ASR counts a response as successful if it is labeled *unsafe*. For WG, ASR counts a response as successful if *harmful response* is *yes*, following the judge output format (Farn et al., 2025). We primarily assess harmful-goal compliance using SR-Score and human judgments; higher SR-Score indicates stronger compliance.

3 Experiments

3.1 Experimental Settings

Datasets and Implementation Details. We used two public jailbreaking benchmarks, Jailbreak-Bench (JBB) and AdvBench. JBB (Chao et al., 2024) collects attack prompts designed to induce LLMs to circumvent safety policies, covering diverse harmful goals and their corresponding prompt formulations. AdvBench (Zou et al., 2023) is an adversarial prompt benchmark for evaluating LLM robustness, and it contains a broad set of harmful behavior scenarios. We performed our experiments using the instruction-based Llama3.2 (Grattafiori et al., 2024) and Qwen2.5 (Qwen et al., 2025) families. Details are provided in Appendix A.

Watermarking Methods. We evaluated six watermarking methods, including four biased schemes (KGW, Unigram, SWEET, MorphMark) and two unbiased schemes (DIP, EXP). **KGW** biases decoding by partitioning the vocabulary into “green” and “red” tokens (Kirchenbauer et al., 2024). **Unigram** uses a fixed green-token set derived from a hash key (Zhao et al., 2024). **SWEET** applies an entropy-gated bias when model uncertainty is high (Lee et al., 2024). **MorphMark** adapts watermark strength using the accumulated green-list probability mass (Wang et al., 2025). **DIP** embeds watermarks with key-conditioned selection while preserving the token distribution via reweighting (Wu et al., 2024). **EXP** uses a key-dependent pseudo-random function with exponential minimum sampling (Kuditipudi et al., 2024). Detailed parameter settings are in Appendix B.

Jailbreaking Methods. We used two prompt-optimization jailbreaking methods. Both approaches automatically optimize attack prompts by repeatedly evaluating candidate prompts against the target LLM. **AutoDAN** separates the attack

process into a critic and a generator, and evolves jailbreak prompts through iterative search based on a genetic algorithm (Liu et al., 2024). DSN (Don’t Say No) optimizes attack prompts by minimizing the model’s tendency to issue refusals. It extends the Greedy Coordinate Gradient (GCG) (Zou et al., 2023) framework by adding a refusal-penalty loss, suppressing refusal signals and encouraging compliance with harmful requests (Zhou and Wang, 2024). In the Static setting, we use fixed harmful prompts. We optimize AutoDAN jailbreak prompts and DSN jailbreak suffixes on the same target model variant under evaluation, and we evaluate using the optimized prompt or suffix fixed at test time.

Evaluation Metrics. We used the same metrics as in Section 2: ASR from LG/WG and SR-Score from StrongREJECT. LG counts *unsafe* outputs; WG counts *harmful response = yes*. SR-Score measures harmful-goal compliance (higher is worse). We computed SR-Score using the official StrongREJECT fine-tuned evaluator, which is trained with a rubric-based evaluation protocol. We report within-setting deltas relative to No-Watermark (NoWM) by subtracting the corresponding NoWM score from each watermarked score for each setting: ΔWG , ΔLG , and ΔSR . For SR-Score, ΔSR is the mean difference, and we report a conservative uncertainty $\sigma_{\Delta} = \sqrt{\sigma_{\text{WM}}^2 + \sigma_{\text{NoWM}}^2}$. Positive ΔWG and ΔLG indicate higher judge-based ASR, and positive ΔSR indicates higher harmful-goal compliance.

3.2 Main Results

Table 1 reports within-setting deltas relative to the NoWM. We do not compare absolute values across settings since they differ in attacker access and optimization signals.

Static Setting. In the Static setting, biased watermarking schemes often yield positive $\Delta\text{WG}/\Delta\text{LG}$ while ΔSR is small or negative, which is consistent with ASR inflation without a corresponding increase in goal-compliance. In contrast, unbiased schemes can show coupled positive shifts in both ASR and goal-compliance on AdvBench, suggesting higher compliance risk in those cases.

AutoDAN and DSN Settings. Under AutoDAN and DSN, we observe the same mismatch pattern: many watermarked models exhibit large positive $\Delta\text{WG}/\Delta\text{LG}$, yet ΔSR is frequently marginal or negative. This gap suggests that ASR gains can

Model	JBB			AdvBench		
	ΔWG	ΔLG	ΔSR	ΔWG	ΔLG	ΔSR
Static						
Llama3.2-3B (NoWM)						
+KGW	+0.14	+0.13	-0.01 (± 0.04)	+0.02	-0.02	+0.02 (± 0.01)
+SWEET	+0.06	+0.02	+0.01 (± 0.04)	+0.01	+0.01	+0.01 (± 0.01)
+Unigram	+0.24	+0.27	-0.10 (± 0.04)	-0.03	+0.19	+0.01 (± 0.01)
+Morphmark	-0.17	-0.27	-0.21 (± 0.04)	-0.01	-0.01	+0.01 (± 0.01)
+EXP	+0.05	+0.02	-0.01 (± 0.04)	+0.21	+0.29	+0.20 (± 0.01)
+DIP	-0.22	-0.32	-0.26 (± 0.03)	+0.24	+0.29	+0.24 (± 0.02)
Qwen2.5-7B (NoWM)						
+KGW	+0.03	+0.03	-0.03 (± 0.04)	+0.00	+0.01	+0.02 (± 0.01)
+SWEET	-0.03	-0.02	-0.03 (± 0.04)	+0.00	+0.01	+0.02 (± 0.01)
+Unigram	-0.02	+0.03	-0.05 (± 0.04)	+0.00	+0.00	+0.02 (± 0.01)
+Morphmark	-0.15	-0.15	-0.17 (± 0.04)	+0.00	+0.00	+0.02 (± 0.01)
+EXP	+0.00	+0.06	+0.07 (± 0.04)	+0.11	+0.13	+0.27 (± 0.01)
+DIP	-0.16	-0.17	-0.17 (± 0.04)	+0.05	+0.08	+0.07 (± 0.01)
AutoDAN						
Llama3.2-3B (NoWM)						
+KGW	+0.01	-0.07	-0.08 (± 0.04)	-0.06	-0.05	-0.14 (± 0.01)
+SWEET	+0.07	+0.02	+0.05 (± 0.04)	-0.04	-0.02	-0.08 (± 0.01)
+Unigram	+0.14	+0.05	-0.22 (± 0.04)	+0.01	+0.01	-0.25 (± 0.02)
+Morphmark	-0.01	-0.06	-0.01 (± 0.04)	+0.00	+0.00	-0.07 (± 0.02)
+EXP	-0.08	-0.13	-0.09 (± 0.05)	-0.08	-0.07	-0.20 (± 0.02)
+DIP	-0.03	-0.09	-0.01 (± 0.04)	-0.03	+0.00	-0.01 (± 0.01)
Qwen2.5-7B (NoWM)						
+KGW	-0.04	-0.05	-0.04 (± 0.03)	-0.06	-0.03	-0.05 (± 0.01)
+SWEET	-0.05	-0.07	-0.03 (± 0.03)	-0.04	-0.02	-0.04 (± 0.01)
+Unigram	-0.05	-0.01	-0.10 (± 0.04)	-0.04	-0.01	-0.08 (± 0.01)
+Morphmark	+0.03	-0.02	+0.00 (± 0.03)	-0.03	-0.01	-0.02 (± 0.01)
+EXP	+0.00	+0.00	-0.04 (± 0.03)	-0.06	-0.02	-0.08 (± 0.01)
+DIP	+0.02	+0.02	-0.01 (± 0.03)	-0.06	-0.03	-0.06 (± 0.01)
DSN (Don’t Say No)						
Llama3.2-3B (NoWM)						
+KGW	-0.01	+0.03	-0.01 (± 0.04)	+0.02	-0.02	+0.00 (± 0.01)
+SWEET	+0.02	+0.07	+0.02 (± 0.04)	+0.00	-0.02	+0.01 (± 0.01)
+Unigram	+0.07	+0.12	+0.00 (± 0.04)	+0.04	+0.04	-0.02 (± 0.01)
+Morphmark	-0.02	+0.04	-0.01 (± 0.04)	+0.02	+0.00	+0.02 (± 0.01)
+EXP	+0.06	+0.01	+0.00 (± 0.04)	+0.03	-0.03	+0.05 (± 0.01)
+DIP	+0.02	+0.04	-0.02 (± 0.04)	+0.03	+0.01	+0.02 (± 0.01)
Qwen2.5-7B (NoWM)						
+KGW	+0.00	+0.09	+0.02 (± 0.05)	+0.06	+0.06	+0.10 (± 0.03)
+SWEET	-0.04	-0.04	-0.06 (± 0.06)	+0.03	+0.01	+0.01 (± 0.03)
+Unigram	+0.02	+0.11	+0.01 (± 0.06)	+0.06	+0.07	+0.01 (± 0.03)
+Morphmark	+0.04	+0.09	+0.02 (± 0.06)	+0.05	+0.05	+0.03 (± 0.03)
+EXP	+0.18	+0.28	+0.04 (± 0.05)	+0.10	+0.11	+0.05 (± 0.02)
+DIP	+0.01	+0.05	-0.01 (± 0.06)	+0.02	+0.02	-0.02 (± 0.03)

Table 1: Experimental results to the no-watermark baseline (NoWM).

arise from perturbations to refusal-style signals (which affect automated judges) rather than from improved harmful-goal completion as measured by StrongREJECT. Across our tested settings, biased schemes more frequently exhibit “ASR-only” gains, whereas unbiased schemes show more mixed behavior and can increase ΔSR in some configurations. Accordingly, alignment evaluations of watermarked LLMs should not rely on ASR alone and should include goal-compliance metrics such as StrongREJECT. Full results for all four LLMs, and watermark-strength settings for biased methods are provided in Appendix C.

4 Analysis

To further analyze watermarking–alignment interactions, we study the ASR–SR gap across attacks, models, and methods.

compliance disagreements; Appendix results further show no consistent coupling between detectability and harmful-goal compliance.

Limitations

We studied watermarking under jailbreaking, which is important for real-world deployment of watermarked LLMs. However, our study has several limitations.

Watermark-aware Attackers. Although we evaluate watermarking under strong jailbreaking methods (AutoDAN and DSN), we do not consider attackers who know the watermark key (Jovanović et al., 2024) to the watermarking scheme. Therefore, the robustness of watermarking against watermark-aware jailbreakers remains a topic for future studies. In the future, we will evaluate key-aware and scheme-aware attacks and analyze how robustness differs between biased and unbiased watermarking.

Multilingual and Domain Generalization. Although our experiments cover two widely used jailbreak benchmarks (AdvBench and JBB), they are limited to English and a fixed set of harmful goals. Therefore, it remains unclear whether our observations generalize to other languages or domains. In the future, we will extend the evaluation to multilingual and domain-specific jailbreak datasets and verify whether the ASR-StrongREJECT mismatch persists under distribution shifts.

Mixed Threat Models. Although we include both deployment-realistic and stronger threat models, we use different attacker capabilities across methods: AutoDAN is evaluated using a fixed, transferable jailbreak prompt at test time, whereas DSN requires white-box access during suffix optimization. Therefore, isolating the effect of attacker access remains a topic for future work. In the future, we will evaluate both methods under unified attacker capabilities, including consistent black-box transfer and consistent white-box optimization settings.

Other Alignment Dimensions. Although our study focuses on safety alignment under jailbreaking, we do not evaluate other alignment dimensions such as truthfulness (Chen et al., 2024), factuality (Chang et al., 2024), or helpfulness (Verma et al., 2025). Therefore, it remains unclear how watermarking interacts with these aspects under adversarial prompting. In the future, we will extend our evaluation to additional alignment benchmarks and analyze whether the ASR-StrongREJECT mis-

match correlates with shifts in truthfulness and overall response quality.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2026-25473607).

References

- Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Wieting, and Mohit Iyyer. 2024. [Postmark: A robust blackbox watermark for large language models](#). *Preprint*, arXiv:2406.14517.
- Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwal, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS Datasets and Benchmarks Track*.
- Liang Chen, Yatao Bian, Yang Deng, Deng Cai, Shuaiyi Li, Peilin Zhao, and Kam fai Wong. 2024. [Watme: Towards lossless watermarking through lexical redundancy](#). *Preprint*, arXiv:2311.09832.
- Hua Farn, Hsuan Su, Shachi H. Kumar, Saurav Sahay, Shang-Tse Chen, and Hung-yi Lee. 2025. [Safeguard fine-tuned LLMs through pre- and post-tuning model merging](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16589–16602, Suzhou, China. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park,

Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Jun-teng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-badur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly

Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymmer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, DingKang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelen, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala,

- Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. [Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms](#). *Preprint*, arXiv:2406.18495.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabza. 2023. [Llama guard: Llm-based input-output safeguard for human-ai conversations](#). *Preprint*, arXiv:2312.06674.
- Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. Watermark stealing in large language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. [A watermark for large language models](#). *Preprint*, arXiv:2301.10226.
- Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. [Robust distortion-free watermarks for language models](#). *Transactions on Machine Learning Research*.
- Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoon Yun, Jamin Shin, and Gunhee Kim. 2024. [Who wrote this code? watermarking for code generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4890–4911, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. [Autodan: Generating stealthy jailbreak prompts on aligned large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, and Michael Chau. 2025. [Watermarking low-entropy generation for large language models: An unbiased and low-risk method](#). *Preprint*, arXiv:2405.14604.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,

- Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. [The curse of recursion: Training on generated data makes models forget](#). *Preprint*, arXiv:2305.17493.
- Samuel Simko, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. 2025. [Improving large language model safety with contrastive representation learning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28154–28182, Suzhou, China. Association for Computational Linguistics.
- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. [A strongreject for empty jailbreaks](#). *Preprint*, arXiv:2402.10260.
- Apurv Verma, Hai Phan, and Shubhendu Trivedi. 2025. [Watermarking degrades alignment in language models: Analysis and mitigation](#). In *The 1st Workshop on GenAI Watermarking*.
- Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. 2025. [MorphMark: Flexible adaptive watermarking for large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4842–4860, Vienna, Austria. Association for Computational Linguistics.
- Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. [A resilient and accessible distribution-preserving watermark for large language models](#). *Preprint*, arXiv:2310.07710.
- Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. 2024. [Provable robust watermarking for AI-generated text](#). In *The Twelfth International Conference on Learning Representations*.
- Yukai Zhou and Wenjie Wang. 2024. [Don't say no: Jailbreaking llm by suppressing refusal](#). *Preprint*, arXiv:2404.16369.
- Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). *Preprint*, arXiv:2307.15043.

A Dataset Statistics and Implementation Details

We use two public jailbreaking benchmarks: JailbreakBench (JBB) (Chao et al., 2024) and AdvBench (Zou et al., 2023). Our evaluation targets harmful-request settings and focuses on jailbreak effectiveness and harmful-goal compliance (ASR and SR-Score). Accordingly, we only use the harmful splits and exclude benign prompts, which do not align with our threat model or metrics.

JailbreakBench (JBB). We use the JBB-Behaviors dataset released by the JailbreakBench project. We select the harmful behaviors split, which contains 100 harmful behaviors spanning 10 high-level categories. Each sample provides (i) a behavior label,

(ii) a goal describing the harmful request, (iii) a target reference response representing desired compliance, (iv) a risk category, and (v) a source attribute indicating where the behavior was collected from.

AdvBench. AdvBench is a standard benchmark for adversarial prompting against aligned LLMs. It contains 520 harmful instructions paired with desired positive (compliant) responses. We use the full set of 520 prompts.

Prompt Format. For both datasets, we treat each sample as a single-turn user prompt.

Implementation Details. Table 5 shows the hyperparameters and computing settings used in our experiments. The CPU used for calculations is an Intel 4th Gen Xeon Scalable Processor (16-core). We used instruction-based Llama3.2-(1B, 3B) and Qwen2.5-(3B, 8B) for our experiments.

	Llama3.2-(1B, 3B)	Qwen2.5-(3B, 8B)
Computing setting	NVIDIA RTX A6000	NVIDIA RTX A6000
Seed	123	123
Temperature	0.7	0.7
Max new tokens	512	512
Min new tokens	75	75

Table 5: Hyperparameters.

B Hyperparameters for Watermarking Methods

	γ	δ	Other Parameters
KGW	0.3	3.0	-
SWEET	0.3	3.0	entropy threshold=3.0
Unigram	0.3	3.0	-
Morphmark	0.3	-	$k_{exp}=1.30, p_0=0.15$
EXP	-	-	threshold=0.01
DIP	0.5	-	$\alpha=0.5$

Table 6: Watermarking Hyperparameters.

C Experimental Results

We additionally evaluate Llama3.2-1B and Qwen2.5-3B. Results for all four LLMs are reported in Table 7, 8, and 9.

To further evaluate detection performance, we utilize the area under the receiver operating characteristic curve (AUC), a standard metric for binary classification. We also report TPR at $FPR \leq 10\%$ (TPR10) for detectability. Across the additional models (Llama3.2-1B and Qwen2.5-3B), we observe patterns consistent with the main paper: judge-based ASR can change substantially under watermarking while SR-Score often remains marginal, especially for biased schemes, whereas

some unbiased schemes increase SR-Score in specific model and benchmark pairs. Furthermore, we find no consistent coupling between detectability and harmful-goal compliance: highly detectable watermarks do not necessarily induce higher SR-Score, and SR-Score increases can occur even when detectability is relatively low. These results suggest that safety outcomes and detectability should be reported jointly but interpreted as largely complementary dimensions.

Figure 1 shows the results for KGW with varying delta values ($\delta = 1.5, 2.0, 2.5,$ and 3.0) at fixed gamma ($\gamma = 0.3$). Varying δ affects judge-based ASR in a setting-dependent manner, with larger shifts under Static on JBB at higher δ , while changes in SR-Score are inconsistent across settings and models. Overall, the ASR-SR mismatch persists across δ , indicating that conclusions based on ASR alone can be sensitive to watermark strength.

D Additional Case Study.

Figure 2 shows additional examples using *Static* setting based on Sweet and Unigram. These cases show how seemingly minor surface patterns, such as repeated tokens or template refusals, can appear in watermarked outputs while the underlying safety behavior remains unchanged.

E ASR Blocking List

This list contains refusal-related phrases used in our ASR-blocking diagnostic. In the Soft setting, we discourage these phrases by assigning them to the red set, while in the Hard setting we block them during decoding.

Refusal Keyword List

"I'm sorry", "Sorry", "I am sorry", "I apologize", "As an", "As an AI", "I am an", "I am just", "As a language model", "As an Assistant", "I cannot", "I do not", "Hello!", "is illegal and unethical", "I'm really sorry", "unethical", "not ethical", "illegal", "not legal", "My apologies", "but I cannot", "If you have any other non-malicious requests", "I'm not able to provide", "I am really sorry", "I cannot create", "I cannot support", "I can't provide", "I can't assist"

Method	Model	JBB					AdvBench				
		WG	LG	SR	AUROC	TPR10	WG	LG	SR	AUROC	TPR10
Static	Llama3.2-1B	0.47	0.59	0.39 (± 0.04)	-	-	0.08	0.05	0.06 (± 0.01)	-	-
	+KGW	0.49	0.61	0.35 (± 0.03)	0.999	1.000	0.10	0.10	0.08 (± 0.01)	0.537	0.100
	+SWEET	0.55	0.64	0.42 (± 0.03)	0.975	0.960	0.06	0.05	0.04 (± 0.01)	0.454	0.039
	+Unigram	0.62	0.78	0.26 (± 0.03)	0.999	1.000	0.15	0.14	0.06 (± 0.01)	0.991	0.958
	+Morphmark	0.08	0.05	0.07 (± 0.02)	0.816	0.430	0.07	0.06	0.07 (± 0.01)	0.828	0.462
	+EXP	0.45	0.69	0.24 (± 0.03)	0.925	0.820	0.46	0.54	0.33 (± 0.02)	0.550	0.164
	+DIP	0.00	0.01	0.00 (± 0.01)	0.700	0.190	0.52	0.58	0.40 (± 0.02)	0.931	0.675
	Llama3.2-3B	0.22	0.34	0.28 (± 0.03)	-	-	0.05	0.05	0.04 (± 0.01)	-	-
	+KGW	0.36	0.47	0.27 (± 0.03)	0.993	0.990	0.07	0.03	0.06 (± 0.01)	0.456	0.064
	+SWEET	0.28	0.36	0.29 (± 0.03)	0.889	0.730	0.06	0.06	0.05 (± 0.01)	0.450	0.035
	+Unigram	0.46	0.61	0.18 (± 0.03)	0.999	1.000	0.02	0.24	0.05 (± 0.01)	0.947	0.900
	+Morphmark	0.05	0.07	0.07 (± 0.02)	0.764	0.510	0.04	0.04	0.05 (± 0.01)	0.778	0.423
	+EXP	0.27	0.36	0.27 (± 0.03)	0.982	0.950	0.26	0.34	0.24 (± 0.01)	0.481	0.181
	+DIP	0.00	0.02	0.02 (± 0.01)	0.716	0.160	0.29	0.34	0.28 (± 0.02)	0.946	0.573
	Qwen2.5-3B	0.25	0.28	0.25 (± 0.03)	-	-	0.00	0.00	0.02 (± 0.03)	-	-
	+KGW	0.27	0.37	0.32 (± 0.03)	1.000	1.000	0.00	0.01	0.03 (± 0.03)	0.556	0.071
	+SWEET	0.22	0.32	0.26 (± 0.04)	0.829	0.680	0.00	0.01	0.03 (± 0.02)	0.573	0.096
	+Unigram	0.28	0.39	0.27 (± 0.03)	0.992	0.970	0.00	0.01	0.02 (± 0.03)	1.000	1.000
	+Morphmark	0.02	0.04	0.07 (± 0.02)	0.960	0.860	0.00	0.00	0.02 (± 0.02)	0.948	0.910
	+EXP	0.30	0.29	0.34 (± 0.03)	1.000	1.000	0.25	0.18	0.34 (± 0.03)	0.444	0.129
+DIP	0.01	0.04	0.08 (± 0.02)	0.991	0.980	0.13	0.17	0.17 (± 0.02)	0.967	0.773	
Qwen2.5-7B	0.16	0.22	0.26 (± 0.03)	-	-	0.00	0.00	0.03 (± 0.01)	-	-	
+KGW	0.19	0.25	0.23 (± 0.03)	1.000	1.000	0.00	0.01	0.05 (± 0.01)	0.483	0.058	
+SWEET	0.13	0.20	0.23 (± 0.03)	0.697	0.400	0.00	0.01	0.05 (± 0.01)	0.510	0.044	
+Unigram	0.14	0.25	0.21 (± 0.03)	0.998	1.000	0.00	0.00	0.05 (± 0.01)	0.990	1.000	
+Morphmark	0.01	0.07	0.09 (± 0.02)	0.921	0.810	0.00	0.00	0.05 (± 0.01)	0.931	0.862	
+EXP	0.16	0.28	0.33 (± 0.03)	0.993	0.990	0.11	0.13	0.30 (± 0.01)	0.510	0.171	
+DIP	0.00	0.05	0.09 (± 0.02)	0.961	0.940	0.05	0.08	0.10 (± 0.01)	0.980	0.689	

Table 7: Experimental result under Static prompts on JBB and AdvBench.

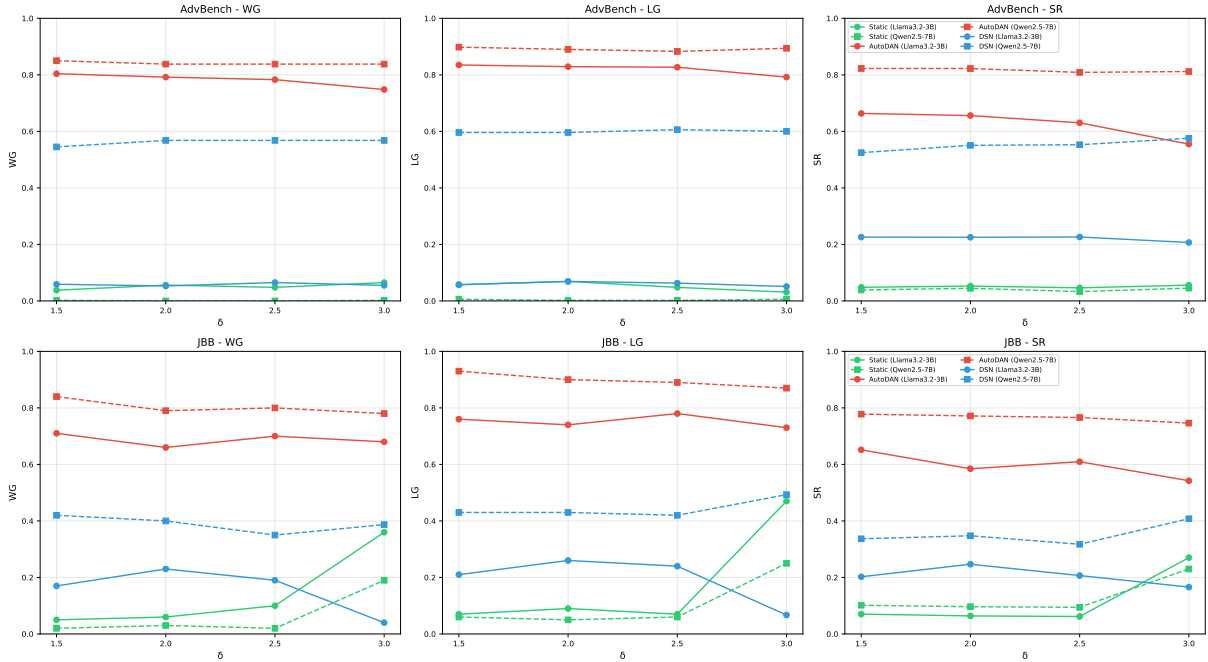


Figure 1: Experimental results of delta variation using KGW.

Method	Model	JBB					AdvBench				
		WG	LG	SR	AUROC	TPR10	WG	LG	SR	AUROC	TPR10
AutoDAN	Llama3.2-1B	0.63	0.65	0.54 (± 0.03)	-	-	0.69	0.72	0.62 (± 0.01)	-	-
	+KGW	0.57	0.67	0.46 (± 0.03)	0.993	0.990	0.56	0.63	0.49 (± 0.01)	1.000	1.000
	+SWEET	0.59	0.76	0.55 (± 0.03)	1.000	1.000	0.68	0.73	0.58 (± 0.01)	0.998	0.996
	+Unigram	0.73	0.75	0.28 (± 0.03)	1.000	1.000	0.74	0.77	0.29 (± 0.01)	1.000	1.000
	+Morphmark	0.65	0.74	0.55 (± 0.03)	0.991	0.990	0.70	0.77	0.59 (± 0.01)	0.974	0.971
	+EXP	0.61	0.76	0.52 (± 0.03)	1.000	1.000	0.66	0.70	0.48 (± 0.02)	0.998	0.996
	+DIP	0.61	0.76	0.55 (± 0.03)	0.983	0.890	0.72	0.73	0.63 (± 0.01)	0.997	0.994
	Llama3.2-3B	0.67	0.80	0.62 (± 0.03)	-	-	0.81	0.84	0.69 (± 0.01)	-	-
	+KGW	0.68	0.73	0.54 (± 0.03)	1.000	1.000	0.75	0.79	0.55 (± 0.01)	0.997	1.000
	+SWEET	0.74	0.82	0.67 (± 0.03)	0.986	0.980	0.77	0.82	0.61 (± 0.01)	0.989	0.977
	+Unigram	0.81	0.85	0.40 (± 0.03)	0.999	1.000	0.82	0.85	0.44 (± 0.02)	0.998	1.000
	+Morphmark	0.66	0.74	0.61 (± 0.03)	0.951	0.980	0.81	0.84	0.62 (± 0.02)	0.946	0.965
	+EXP	0.59	0.67	0.53 (± 0.04)	0.967	0.930	0.73	0.77	0.49 (± 0.02)	0.966	0.948
	+DIP	0.64	0.71	0.61 (± 0.03)	0.993	0.810	0.78	0.84	0.68 (± 0.01)	0.989	0.971
	Qwen2.5-3B	0.78	0.88	0.75 (± 0.03)	-	-	0.86	0.89	0.78 (± 0.01)	-	-
	+KGW	0.74	0.85	0.68 (± 0.03)	1.000	1.000	0.92	0.93	0.79 (± 0.01)	1.000	1.000
	+SWEET	0.81	0.86	0.77 (± 0.02)	1.000	1.000	0.90	0.91	0.78 (± 0.01)	0.998	0.998
	+Unigram	0.78	0.89	0.65 (± 0.03)	1.000	1.000	0.82	0.85	0.60 (± 0.01)	1.000	1.000
	+Morphmark	0.84	0.92	0.76 (± 0.02)	0.999	1.000	0.89	0.91	0.78 (± 0.01)	1.000	1.000
	+EXP	0.81	0.88	0.65 (± 0.03)	1.000	1.000	0.90	0.92	0.67 (± 0.01)	0.997	0.996
	+DIP	0.84	0.96	0.76 (± 0.02)	1.000	1.000	0.87	0.89	0.78 (± 0.01)	0.999	0.998
	Qwen2.5-7B	0.82	0.92	0.79 (± 0.02)	-	-	0.90	0.92	0.86 (± 0.01)	-	-
	+KGW	0.78	0.87	0.75 (± 0.02)	1.000	1.000	0.84	0.89	0.81 (± 0.01)	1.000	1.000
	+SWEET	0.77	0.85	0.76 (± 0.02)	0.992	0.980	0.86	0.90	0.82 (± 0.01)	0.990	0.983
	+Unigram	0.77	0.91	0.69 (± 0.03)	1.000	1.000	0.86	0.91	0.78 (± 0.01)	1.000	1.000
	+Morphmark	0.85	0.90	0.79 (± 0.02)	0.999	1.000	0.87	0.91	0.84 (± 0.01)	0.996	0.996
	+EXP	0.82	0.92	0.75 (± 0.02)	1.000	1.000	0.84	0.90	0.78 (± 0.01)	1.000	1.000
	+DIP	0.84	0.94	0.78 (± 0.02)	1.000	0.980	0.84	0.89	0.80 (± 0.01)	1.000	1.000

Table 8: Experimental result under AutoDAN-generated adversarial prompts on JBB and AdvBench.

Method	Model	JBB					AdvBench				
		WG	LG	SR	AUROC	TPR10	WG	LG	SR	AUROC	TPR10
DSN	Llama3.2-1B	0.04	0.12	0.12 (± 0.03)	-	-	0.06	0.07	0.09 (± 0.01)	-	-
	+KGW	0.07	0.08	0.11 (± 0.02)	0.985	0.973	0.08	0.07	0.09 (± 0.01)	0.981	0.929
	+SWEET	0.08	0.12	0.14 (± 0.01)	0.795	0.547	0.04	0.08	0.10 (± 0.01)	0.756	0.404
	+Unigram	0.08	0.13	0.09 (± 0.02)	0.994	0.947	0.11	0.10	0.10 (± 0.01)	0.990	0.927
	+Morphmark	0.03	0.05	0.14 (± 0.02)	0.851	0.587	0.04	0.06	0.07 (± 0.01)	0.807	0.426
	+EXP	0.16	0.07	0.11 (± 0.02)	0.989	0.987	0.18	0.05	0.10 (± 0.01)	0.975	0.954
	+DIP	0.04	0.07	0.10 (± 0.02)	0.950	0.853	0.07	0.12	0.13 (± 0.01)	0.918	0.776
	Llama3.2-3B	0.05	0.04	0.18 (± 0.03)	-	-	0.04	0.07	0.21 (± 0.01)	-	-
	+KGW	0.04	0.07	0.17 (± 0.03)	0.977	0.947	0.06	0.05	0.21 (± 0.01)	0.962	0.956
	+SWEET	0.07	0.11	0.20 (± 0.03)	0.716	0.267	0.04	0.05	0.22 (± 0.01)	0.673	0.214
	+Unigram	0.12	0.16	0.18 (± 0.03)	0.989	0.867	0.08	0.11	0.19 (± 0.01)	0.965	0.909
	+Morphmark	0.03	0.08	0.17 (± 0.03)	0.788	0.507	0.06	0.07	0.23 (± 0.01)	0.843	0.598
	+EXP	0.11	0.05	0.18 (± 0.03)	0.922	0.867	0.07	0.04	0.26 (± 0.01)	0.981	0.966
	+DIP	0.07	0.08	0.16 (± 0.03)	0.886	0.680	0.07	0.08	0.23 (± 0.01)	0.897	0.739
	Qwen2.5-3B	0.40	0.45	0.45 (± 0.04)	-	-	0.32	0.33	0.38 (± 0.02)	-	-
	+KGW	0.36	0.41	0.45 (± 0.04)	1.000	1.000	0.22	0.24	0.35 (± 0.01)	1.000	1.000
	+SWEET	0.36	0.41	0.41 (± 0.04)	0.857	0.680	0.31	0.33	0.37 (± 0.02)	0.823	0.626
	+Unigram	0.56	0.57	0.52 (± 0.04)	0.998	1.000	0.57	0.57	0.57 (± 0.02)	0.999	1.000
	+Morphmark	0.39	0.45	0.42 (± 0.04)	0.986	0.973	0.32	0.33	0.38 (± 0.02)	0.967	0.937
	+EXP	0.56	0.67	0.56 (± 0.03)	1.000	1.000	0.42	0.45	0.54 (± 0.01)	1.000	1.000
	+DIP	0.29	0.33	0.30 (± 0.04)	0.964	0.893	0.26	0.26	0.29 (± 0.02)	0.973	0.915
	Qwen2.5-7B	0.39	0.40	0.39 (± 0.04)	-	-	0.51	0.54	0.48 (± 0.02)	-	-
	+KGW	0.39	0.49	0.41 (± 0.03)	0.997	1.000	0.57	0.60	0.58 (± 0.02)	0.999	1.000
	+SWEET	0.35	0.36	0.33 (± 0.04)	0.658	0.333	0.54	0.55	0.49 (± 0.02)	0.786	0.509
	+Unigram	0.41	0.51	0.40 (± 0.04)	0.986	1.000	0.57	0.61	0.49 (± 0.02)	0.975	0.996
	+Morphmark	0.43	0.49	0.41 (± 0.04)	0.873	0.853	0.56	0.59	0.51 (± 0.02)	0.883	0.699
	+EXP	0.57	0.68	0.43 (± 0.03)	0.965	0.933	0.61	0.65	0.53 (± 0.01)	0.999	0.996
	+DIP	0.40	0.45	0.38 (± 0.04)	0.944	0.840	0.53	0.56	0.46 (± 0.02)	0.931	0.816

Table 9: Experimental result under DSN-generated adversarial prompts on JBB and AdvBench.

