

Implicit Unwatermarking in Pseudo-Reference Reranking for LLM Watermarking

Sungwoo Han¹, *Jingun Kwon¹,
Hidetaka Kamigaito², and Manabu Okumura³

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

³Institute of Science Tokyo

77sungwhan@o.cnu.ac.kr

jingun.kwon@cnu.ac.kr

kamigaito.h@is.naist.jp

oku@pi.titech.ac.jp

Abstract

Text watermarking for large language models (LLMs) requires balancing output quality and detection reliability. We study a practically relevant reranking setting in which multiple watermarked outputs are compared against unwatermarked pseudo-references. We show that quality-only reranking can weaken watermark signals by favoring candidates closer to unwatermarked references, effectively acting as implicit unwatermarking. We propose **constrained Minimum Bayes Risk (MBR) selection**, which treats detection as a constraint rather than a weighted objective: it filters out weakly watermarked candidates and then selects the highest-quality remaining output. Experiments across watermarking methods, datasets, and models show that constrained selection substantially improves detectability over quality-only reranking while remaining competitive in task quality. It also achieves a more favorable matched-detection tradeoff than normalized weighted MBR.

1 Introduction

Text watermarking has emerged as a promising approach for attributing LLM-generated text to its source (Kirchenbauer et al., 2023; Liu et al., 2024). In practice, however, watermarking must balance output quality and detection reliability. Stronger watermark signals are easier to detect but can harm generation quality, whereas weaker signals better preserve quality but are harder to detect (Giboulot and Furon, 2024; Wu et al., 2024).

A natural way to improve the quality of watermarked generation is to sample multiple candidates and rerank them before release. Although Minimum Bayes Risk (MBR) decoding and related reranking methods are widely used to improve generation quality (Kumar and Byrne, 2004; Freitag

et al., 2022; Fernandes et al., 2022), their implications for watermark detection remain underexplored. We study a reranking setting in which multiple watermarked candidates are compared against unwatermarked pseudo-references, which serve as a cleaner semantic proxy for quality by avoiding watermark contamination. Related MBR variants have also explored using generated quasi-sources or support hypotheses for reranking (Lyu et al., 2025).

This design introduces a practical failure mode: because the pseudo-references are unwatermarked, candidates with weaker watermark signals can receive higher reranking scores by being more similar to them. Prior work has mainly studied explicit attacks on text watermarks, such as paraphrasing or rewriting, token substitution, scrubbing, and reverse engineering (Krishna et al., 2023; Jovanović et al., 2024; Wu and Chandrasekaran, 2024; Chen et al., 2025). Recent work has also examined watermark detection under jailbreak attacks (Han et al., 2026). WaterSearch (Lin et al., 2026), by contrast, incorporates detection into candidate selection through a weighted objective. Here, we focus on a post-hoc reranking setting in which quality-only selection against unwatermarked pseudo-references can implicitly weaken watermark signals.

In our setting, scalarization can be brittle because quality and detection play different operational roles. Quality should be optimized among acceptable outputs, whereas detection determines whether an output is acceptable at all. This motivates a **constrained MBR formulation** for selection. Because the watermark generator knows the secret key (Kirchenbauer et al., 2023), it can compute a detection score for each candidate before selection. Our method therefore first filters out weakly watermarked candidates and then selects the highest-quality remaining output.

Experiments across watermarking methods,

* Corresponding author

datasets, and models show that the constrained formulation consistently improves the quality-detection tradeoff over quality-only reranking. In the GSM8K analysis, it also outperforms normalized weighted MBR under matched detection targets. Our contribution is not a new watermark injection mechanism. Instead, we identify implicit unwatermarking as a practical failure mode of unwatermarked pseudo-reference reranking and show that treating detection as a constraint is better suited to this setting than scalarized reranking.

2 Method

We consider a multi-candidate watermarking setting in which, for each input, the model generates a set of watermarked candidates together with a set of unwatermarked pseudo-references. The pseudo-references are used only for reranking, while the final selected output remains watermarked. This setting is practically relevant when pseudo-references are used only as an internal quality signal for post-hoc selection: leaving them unwatermarked preserves a cleaner semantic proxy and avoids watermark contamination in reranking.

KGW-based Watermarking Setup. Following KGW-style watermarking, at each decoding step t , the vocabulary is pseudorandomly partitioned using a secret key into a *green list* $\mathcal{G}_t$ and its complement, and the sampler is biased toward tokens in $\mathcal{G}_t$ (Kirchenbauer et al., 2023; Lee et al., 2024; Zhao et al., 2024). As a result, watermarked text tends to contain more green-list tokens than expected under the null hypothesis of no watermark. Given a generated sequence $y = (w_1, \dots, w_T)$ and the secret key, the detector reconstructs $\mathcal{G}_t$ at each step and computes

$$z(y) = \frac{\sum_{t=1}^T \mathbf{1}[w_t \in \mathcal{G}_t] - \gamma T}{\sqrt{T\gamma(1-\gamma)}}, \quad (1)$$

where γ is the expected green-list fraction under the null, and larger values indicate stronger evidence of watermark presence. In our experiments, KGW, SWEET, and MorphMark all use detectors from this excess-green-token family, with scheme-specific watermarking details described in Appendix A.

Minimum Bayes Risk (MBR) Decoding. Given a candidate set $\mathcal{Y} = \{y_1, \dots, y_K\}$ and pseudo-reference set $\mathcal{R} = \{y'_1, \dots, y'_M\}$, we define the

quality score of a candidate as

$$Q(y_i) = \frac{1}{M} \sum_{j=1}^M u(y_i, y'_j), \quad (2)$$

where u is a pseudo-reference-based utility function, such as BERTScore (Zhang et al., 2020).

When the pseudo-references are unwatermarked, quality-only reranking can favor candidates with weaker watermark signals, because such candidates can be more similar to the pseudo-references. A direct fix is weighted MBR:

$$y^* = \arg \max_{y_i \in \mathcal{Y}} (1 - \lambda)Q(y_i) + \lambda D(y_i), \quad (3)$$

where $\lambda \in [0, 1]$ and $D(y_i) = z(y_i)$ is the detection score used only for candidate ranking. In practice, however, this scalarized objective can be brittle because Q and D differ in both scale and operational meaning, so the tradeoff parameter must simultaneously calibrate heterogeneous signals and encode a minimum detection requirement.

Our Constrained MBR Selection. We instead treat detection as a constraint. This matches a deployment perspective in which detectability is a minimum requirement and quality is optimized only among feasible candidates. In other words, detection determines which outputs are acceptable, while quality determines which acceptable output should be selected.

$$y^* = \arg \max_{y_i \in \mathcal{Y}} Q(y_i) \quad \text{s.t.} \quad D(y_i) \geq \tau, \quad (4)$$

where τ is a minimum detection threshold.

Rather than fixing τ , we parameterize the constraint by a keep ratio $r \in [0, 1]$. Let $k = \max(1, \lceil Kr \rceil)$ and let $D_{(k)}$ denote the k -th largest detection score in $\mathcal{Y}$. We define the feasible set as

$$\mathcal{F}_r = \{y_i \in \mathcal{Y} : D(y_i) \geq D_{(k)}\}, \quad (5)$$

and select $y^* = \arg \max_{y_i \in \mathcal{F}_r} Q(y_i)$.

Smaller values of r impose a stricter constraint by retaining fewer candidates, while larger values relax the constraint and preserve more freedom for quality optimization. When $r = 0$, we keep the top-detection candidate, up to ties.

This constrained formulation is designed to counteract implicit unwatermarking by removing weakly watermarked candidates before optimizing quality within the retained set.

WM	Model	Method	CNN/DM				XSum				GovReport				GSM8K		
			R-2↑	BS↑	TPR↑	AUC↑	R-2↑	BS↑	TPR↑	AUC↑	R-2↑	BS↑	TPR↑	AUC↑	Acc↑	TPR↑	AUC↑
KGW	Llama-3.1-8B	No WM	12.1	86.8	0.6	51.2	7.4	87.7	0.4	49.2	13.4	84.1	1.5	51.3	76.4	1.0	49.3
		Baseline	11.7	86.7	73.8	98.2	7.1	87.6	36.4	91.7	12.7	84.2	87.5	99.0	72.0	48.0	91.9
		Q-only MBR	12.5	87.0	60.4	96.7	7.6	88.0	26.0	87.8	13.5	84.3	76.5	98.4	83.4	31.0	88.1
		Ours	12.3	86.9	85.8	99.3	7.4	87.9	44.4	95.6	13.4	84.3	93.5	99.7	81.0	50.6	95.5
		Ours ($\tau=1.96$)	11.3	86.7	72.4	99.3	6.8	87.6	67.0	99.1	11.7	84.1	80.0	99.3	70.8	71.6	99.0
	Qwen3-1.7B	No WM	11.7	86.6	2.4	50.9	4.4	86.9	1.2	51.1	10.2	84.9	4.0	54.5	78.8	0.8	49.6
		Baseline	11.6	86.6	33.4	90.1	4.3	86.8	10.6	77.8	10.0	84.9	24.5	87.4	78.8	6.4	70.8
		Q-only MBR	12.1	86.7	25.6	86.2	4.5	87.0	4.8	72.4	10.2	85.0	17.5	85.7	80.0	6.8	68.3
		Ours	11.4	86.6	77.0	98.6	4.3	86.9	29.6	94.6	9.7	85.0	69.0	98.5	78.0	26.6	90.6
		Ours ($\tau=1.96$)	11.1	86.6	37.2	97.2	4.2	86.9	25.8	95.4	9.5	85.0	29.2	97.3	78.6	5.8	88.4
SWEET	Llama-3.1-8B	Baseline	11.8	86.7	49.4	93.5	7.4	87.8	23.0	83.9	13.1	84.3	62.0	96.8	76.8	35.0	86.0
		Q-only MBR	12.7	87.0	39.6	92.5	7.9	88.0	14.2	78.3	14.0	84.3	55.5	95.5	85.6	17.6	82.2
		Ours	12.2	86.9	80.4	99.1	7.7	87.9	44.4	95.5	13.4	84.3	94.0	99.8	82.4	57.2	96.6
		Ours ($\tau=1.96$)	11.3	86.7	52.8	98.6	6.8	87.8	44.8	97.8	12.0	84.3	62.4	98.6	74.4	73.0	99.0
	Qwen3-1.7B	Baseline	11.9	86.7	11.2	74.3	4.4	86.9	2.6	61.8	10.2	84.9	5.0	70.6	81.4	4.2	58.0
		Q-only MBR	12.1	86.8	8.0	71.5	4.6	87.0	3.6	59.9	10.4	85.0	4.0	65.6	80.6	0.6	57.8
		Ours	11.6	86.5	46.2	96.1	4.6	86.9	24.0	90.6	9.9	84.8	41.0	95.4	78.0	13.2	90.6
		Ours ($\tau=1.96$)	11.2	86.6	35.4	95.4	4.2	86.9	13.6	89.1	9.4	84.8	30.4	94.4	80.8	18.8	89.7
MorphMark	Llama-3.1-8B	Baseline	12.3	86.8	44.0	93.3	7.4	87.7	13.2	82.1	13.4	84.3	31.0	88.5	76.0	13.0	82.2
		Q-only MBR	13.0	87.1	31.0	89.4	7.9	88.0	7.6	77.8	12.7	84.0	21.0	87.9	83.0	9.4	78.0
		Ours	12.3	86.9	82.0	99.2	7.6	87.9	41.4	96.6	13.0	84.2	74.5	99.1	76.0	35.0	95.4
		Ours ($\tau=1.96$)	12.1	87.0	49.0	98.7	7.1	87.9	31.8	97.8	12.1	84.0	32.0	97.3	73.6	29.0	97.3
	Qwen3-1.7B	Baseline	11.6	86.6	13.2	80.1	4.4	86.9	5.0	68.6	10.1	85.0	9.5	75.3	80.2	5.4	65.7
		Q-only MBR	12.3	86.8	8.6	77.5	4.6	87.0	3.2	64.9	10.0	84.9	7.0	71.1	80.0	4.8	62.3
		Ours	11.6	86.6	50.4	96.4	4.5	86.9	31.8	92.5	10.0	84.8	39.5	95.5	78.2	22.2	88.9
		Ours ($\tau=1.96$)	11.4	86.7	14.0	95.0	4.4	87.0	14.6	92.2	9.3	84.9	10.0	94.2	79.0	6.4	86.4

Table 1: Main results. **No WM** denotes the unwatermarked reference point. Acc, TPR, and AUC are in %. Best task quality among watermarked methods within each watermark-model group is shown in **bold**.

Method	KGW		SWEET	
	P@1↑	AUC↑	P@1↑	AUC↑
No WM	19.5	50.2	17.7	51.0
Baseline	11.0	94.1	18.3	80.7
Q-only MBR	17.1	78.5	22.6	56.3
Ours	18.3	95.2	18.3	96.0
Ours ($\tau=1.96$)	14.6	99.4	23.8	91.9

Table 2: Code generation results on HumanEval.

3 Experiments

Datasets and Evaluation Metrics. We evaluate watermarking methods on summarization, mathematical reasoning, and code generation. For summarization, we use CNN/DailyMail (Nallapati et al., 2016), XSum (Narayan et al., 2018), and GovReport (Huang et al., 2021), and report ROUGE-2 (R-2) (Lin, 2004) and BERTScore (BS) against gold references. For mathematical reasoning, we use GSM8K (Cobbe et al., 2021) and report accuracy (Acc). For code generation, we use HumanEval (Chen et al., 2021) and report pass@1 (P@1). For detection, we report AUROC (AUC) and TPR at FPR $\leq 1\%$.

Baselines. We compare three watermarking methods: **KGW** (Kirchenbauer et al., 2023), **SWEET** (Lee et al., 2024), and **MorphMark** (Wang et al., 2025). All three use the same z-score-based detection framework in Eq. 1. In the unwatermarked pseudo-reference reranking setting, we compare **Baseline** single-sample decoding, **quality-only MBR (Q-only MBR)**, and **Con-**

strained MBR (Ours), which filters candidates by detection before selecting the highest-quality retained output. Baseline uses a single watermarked sample without reranking, whereas quality-only MBR and constrained MBR select from a shared multi-sample candidate pool. We additionally compare against **weighted MBR** in the analysis section. Because weighted MBR is defined by a continuous tradeoff parameter rather than a single canonical operating point, we evaluate it through an operating-point sweep rather than reporting one arbitrarily chosen point.

Models and Implementation Details. We use Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-1.7B (Yang et al., 2025), and StarCoder (Li et al., 2023). For each input, we generate $K = 16$ watermarked candidates and $M = 4$ unwatermarked pseudo-references from the same model, and all methods use the same candidate and pseudo-reference sets. We use BERTScore in Eq. 2 for reranking only, while final performance is measured with task-specific metrics. For constrained MBR, we sweep $r \in \{0.0, 0.1, \dots, 1.0\}$ and select the *largest* value that satisfies a prespecified target AUROC requirement for each watermark-model pair. To address concerns about oracle-like tuning of the keep ratio, we also evaluated a deployment-oriented variant that uses a fixed statistical significance threshold, $\tau=1.96$, corresponding to the conventional one-sided $p < 0.05$ criterion under the

null hypothesis. Within each task/model setting, all methods share the same candidate and pseudo-reference sets. Additional implementation details are provided in the Appendix A.

Main Results. Table 1 shows a consistent pattern in the unwatermarked pseudo-reference reranking setting. Quality-only MBR improves task quality but often weakens watermark detection, consistent with implicit unwatermarking. In contrast, constrained selection substantially improves detectability while remaining competitive in task quality. Table 2 shows a similar trend on HumanEval: quality-only reranking improves pass@1 but weakens detection, whereas constrained reranking recovers strong detection performance while retaining much of the quality benefit. These results suggest that the main failure mode is not candidate generation itself, but the interaction between unwatermarked pseudo-references and quality-only selection. Even when all candidate sets are generated under the same watermarking scheme, the reranking rule can materially change the final detection strength.

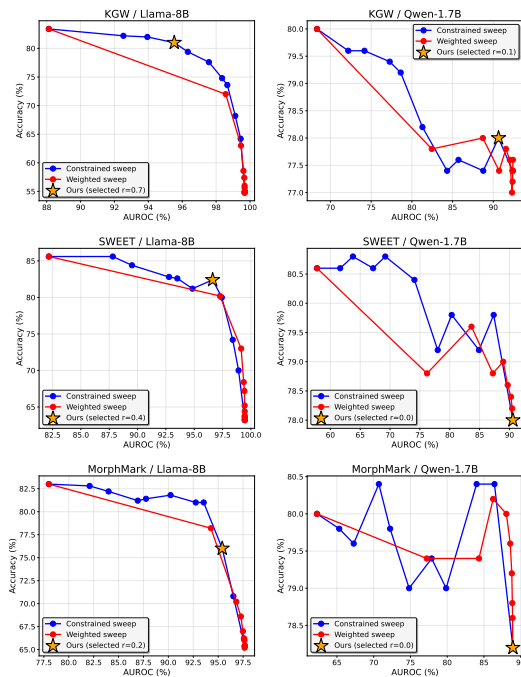


Figure 1: Accuracy-AUROC trade-offs on GSM8K for weighted MBR and constrained selection. Blue points trace different keep ratios r for constrained selection, and red points trace different λ values for weighted MBR. The star marks the selected constrained operating point. On Qwen3-1.7B, weaker watermark separability leaves limited room for differentiation.

4 Analysis

Comparison to Weighted MBR. We compare constrained selection against weighted MBR on GSM8K, where accuracy is only weakly aligned with the BERTScore proxy used for reranking. For weighted MBR, we normalize quality and detection scores within each candidate set and sweep λ . For each watermark-model pair, we fix our operating point using the selected keep ratio r , then choose the smallest λ in Eq. 3 whose AUROC matches or exceeds ours; when no exact match exists, we use the nearest weighted operating point that meets or exceeds the constrained AUROC. Figure 1 shows that constrained selection achieves a more favorable accuracy-AUROC trade-off than weighted MBR under matched detection targets, and also includes the detection-only extreme ($\lambda=1$ or $r=0$).

Computational Cost. Our approach has essentially the same computational structure as standard MBR decoding (Kumar and Byrne, 2004). Both generate K candidates and M pseudo-references, and compute $K \times M$ BERTScore evaluations for reranking. Constrained selection adds only a lightweight z-score computation, with no additional neural inference. Additional effects of sample size, LLM-based judgment, robustness results, output lengths, case studies, and candidate-level quality-detection correlations are provided in Appendices B to G.

When can constrained selection be preferable? Constrained selection can be preferable when detectability acts as a minimum acceptability requirement rather than as a continuously tradable utility. This setting is relevant for watermark deployment: once an output satisfies a desired detection floor, further increasing its detection score need not compensate for a loss in task quality. The distinction is especially important when pseudo-reference quality Q and detection score D are negatively correlated, since quality-only reranking can then favor weakly watermarked candidates. In contrast to weighted scalarization, constrained selection first restricts the candidate pool to a feasible set $\mathcal{F}_r$ and then optimizes quality only within that set. Appendix H gives a pairwise condition showing how weighted scalarization can select a lower-quality candidate even when both candidates already satisfy the detection floor.

5 Conclusion

We identified *implicit unwatermarking* as a practical failure mode of reranking watermarked candidates against unwatermarked pseudo-references. Across watermarking methods, datasets, and models, treating detection as a constraint rather than a weighted objective improved detectability over quality-only reranking while remaining competitive in task quality. In reranking settings with unwatermarked pseudo-references, watermark-aware selection should treat detectability as a feasibility requirement rather than merely another scalarized objective.

Limitations

We evaluate only logit-based watermarking methods with a z-score-style detection interface, so we do not verify the proposed formulation beyond this family. We also focus on a specific reranking setup with unwatermarked pseudo-references and BERTScore, and do not test alternative pseudo-reference constructions or reranking utilities. In addition, because z-score-based detection depends on sequence length, output length is a potential confound in comparing reranking strategies. Finally, our analysis uses relatively small candidate pools and a fixed pseudo-reference budget, so we do not fully characterize larger-scale regimes. Although unwatermarked pseudo-references are a natural internal quality signal in practice, they are not the only reasonable design choice. Our study focuses on whether a simple constrained formulation is sufficient to address implicit unwatermarking, rather than on developing more elaborate reranking strategies.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2026-25569031, Development of an AIBOM-Based Platform for Strengthening the Reliability and Security of the AI Supply Chain). This work was supported by the Korea Basic Science Institute (KBSI) Young Researcher Infrastructure Support Program funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2026-25473607). Following are results of a study on the “Convergence and Open Sharing System” Project, supported by the Ministry of Education and National Research Foundation of Korea.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone.](#)
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebggen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code.](#)
- Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang.

2025. [De-mark: Watermark removal in large language models](#). In *Forty-second International Conference on Machine Learning*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#).
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. [Quality-aware decoding for neural machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. [High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics](#). *Transactions of the Association for Computational Linguistics*, 10:811–825.
- Eva Giboulot and Teddy Furon. 2024. [Watermax: breaking the LLM watermark detectability-robustness-quality trade-off](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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 Allonsius, 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, Junteng 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 Lakhotia, 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 Çelebi, 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 Raparthy, 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, Wenxin 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 Paula, 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, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan 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, Filippas 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 Khan-delwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, 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 Khabza, 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 Ratanandani, 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, Yuze He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#).
- Sungwoo Han, Sangjun Moon, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2026. [Measuring watermarking under jailbreaking: ASR inflation and goal-compliance mismatch](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 36071–36083, San Diego, California, United States. Association for Computational Linguistics.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. [Efficient attentions for long document summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.
- Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. [Watermark stealing in large language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22570–22593. PMLR.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. [A watermark for large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.
- 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](#).
- Shankar Kumar and William Byrne. 2004. [Minimum Bayes-risk decoding for statistical machine translation](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo 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.

- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. [Starcoder: may the source be with you!](#)
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yukang Lin, Jiahao Shao, Shuoran Jiang, Wentao Zhu, Bingjie Lu, Xiangping Wu, Joanna Siebert, and Qingcai Chen. 2026. [Watersearch: Exploring seed pooling for improving the quality-detectability trade-off in llm watermarking](#).
- Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024. [A survey of text watermarking in the era of large language models](#). *ACM Comput. Surv.*, 57(2).
- Boxuan Lyu, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2025. [Unveiling the power of source: Source-based minimum Bayes risk decoding for neural machine translation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2976–2994, Vienna, Austria. Association for Computational Linguistics.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. 2024. [MarkLLM: An open-source toolkit for LLM watermarking](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 61–71, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Qilong Wu and Varun Chandrasekaran. 2024. [Bypassing LLM watermarks with color-aware substitutions](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8549–8581, Bangkok, Thailand. 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. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#).
- 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*.

A Additional Implementation Details and Experiments

We use MarkLLM, an open-source toolkit for LLM watermarking (Pan et al., 2024), and carefully implement all baselines. For evaluation, we use

WM	Model	Method	CNN/DM				XSum				GSM8K		
			R-2↑	BS↑	TPR↑	AUC↑	R-2↑	BS↑	TPR↑	AUC↑	Acc↑	TPR↑	AUC↑
KGW	Llama-3.2-3B	No WM	10.7	86.4	1.6	50.0	6.1	87.3	0.6	49.4	69.4	1.0	50.9
		Baseline	10.1	86.3	88.6	99.2	5.8	87.2	49.8	95.2	64.8	25.0	85.8
		Q-only MBR	11.2	86.6	82.8	98.7	6.3	87.6	36.4	93.1	76.2	26.2	85.4
		Ours	10.4	86.5	100.0	100.0	5.7	87.4	81.0	99.3	67.6	59.8	96.0
	Qwen3-4B	No WM	11.2	86.2	0.8	49.0	5.4	87.1	1.4	50.8	63.4	1.0	49.0
		Baseline	11.0	86.2	17.6	83.5	5.6	87.2	5.8	68.6	65.8	6.0	66.7
Q-only MBR		11.2	86.3	16.0	81.5	5.6	87.3	6.4	64.6	67.8	3.4	63.1	
Ours		11.0	86.2	60.8	97.4	5.3	87.2	50.2	93.3	64.4	21.2	88.1	
SWEET	Llama-3.2-3B	Baseline	10.3	86.2	81.0	98.3	5.8	87.2	30.4	90.3	67.6	23.2	82.2
		Q-only MBR	11.3	86.6	67.4	96.5	6.8	87.6	25.2	86.2	77.2	10.2	75.5
		Ours	10.7	86.5	98.0	99.9	6.3	87.5	71.6	98.3	69.0	39.6	95.7
	Qwen3-4B	Baseline	11.2	86.2	13.4	78.4	5.6	87.1	1.8	62.3	64.4	3.4	58.3
		Q-only MBR	11.5	86.3	8.4	73.3	5.6	87.3	3.8	59.8	66.0	3.2	56.6
		Ours	10.9	86.1	52.8	96.3	5.4	87.2	19.0	88.7	63.2	25.8	88.5
MorphMark	Llama-3.2-3B	Baseline	11.1	86.5	62.6	97.0	6.1	87.3	28.4	88.5	70.0	13.6	77.5
		Q-only MBR	11.7	86.7	50.0	94.9	6.5	87.6	19.4	84.8	74.8	8.4	76.6
		Ours	10.9	86.5	97.4	99.9	6.1	87.5	80.4	99.2	66.8	42.8	95.1
	Qwen3-4B	Baseline	11.8	86.3	13.6	79.5	5.4	87.2	4.0	65.4	46.0	2.0	60.6
		Q-only MBR	12.1	86.4	12.6	77.0	5.5	87.2	2.0	62.7	47.8	1.8	58.8
		Ours	11.5	86.3	53.2	96.3	5.4	87.2	25.4	91.5	49.0	12.4	84.0

Table 3: Additional experimental results.

500 samples from CNN/DailyMail, XSum, and GSM8K, and 250 samples from GovReport. For HumanEval, we use the full dataset of 164 problems.

Unless otherwise specified, we use sampling with `do_sample=True`, `top_k=50`, `top_p=1.0`, and `temperature=1.0`. We set `max_new_tokens` to 1,000 for GovReport and 200 for CNN/DailyMail, XSum, GSM8K, and HumanEval throughout all experiments.

KGW, SWEET, and MorphMark share a green-list ratio $\gamma = 0.5$, hash key $h = 15485863$, and a context window size of 1. KGW and SWEET use a fixed logit bias $\delta = 2.0$, while SWEET further applies an entropy threshold of 0.9 to watermark only high-entropy tokens. MorphMark adaptively adjusts δ based on token entropy via an exponential function with $k_{\text{exp}} = 1.30$.

For constrained MBR, we prespecified model-specific target AUROC values before the main comparisons. We then sweep $r \in \{0.0, 0.1, \dots, 1.0\}$ and select the *largest* value that satisfies the target for each watermark-model pair. The selected operating point is reused unchanged across all reported datasets rather than re-tuned separately for each dataset. We use prespecified target AUROCs of 0.95 for Llama-3.1-8B and 0.90 for Qwen3-1.7B. We use a lower target for Qwen3-1.7B because watermark separability is substantially weaker for this model in our setup, making the higher target unattainable for some watermark schemes without collapsing the feasible set. The selected values

are $r = 0.7, 0.4, 0.2$ for KGW, SWEET, and MorphMark on Llama-3.1-8B, and $r = 0.1, 0.0, 0.0$ on Qwen3-1.7B. Because a single operating point is selected for each watermark-model pair and then reported separately across datasets, individual per-dataset AUC values in Table 1 may fall slightly below the target threshold.

Table 3 reports additional results on CNN/DM, XSum, and GSM8K using Llama-3.2-3B and Qwen3-4B. The overall pattern matches the main experiments: quality-only MBR often improves task quality while weakening detection, whereas constrained selection substantially strengthens detectability while remaining competitive in quality.

B Effect of Sample Size

Figure 2 shows that n_{ref} has limited effect once n_{cand} is fixed, because detection filtering often leaves only a small feasible set for reranking. By contrast, n_{cand} directly changes the candidate pool and has a much larger effect on the quality-detection tradeoff. Increasing n_{cand} consistently improves detection, but tends to reduce accuracy after a small peak.

C LLM Judgment Evaluation

We also conducted LLM-based evaluation. We use Gemini 2.0 Flash as the LLM judge¹. We sampled 50 instances from CNN/DM, and for each instance, an LLM judge rated the outputs on a 1–5

¹<https://gemini.google.com/>

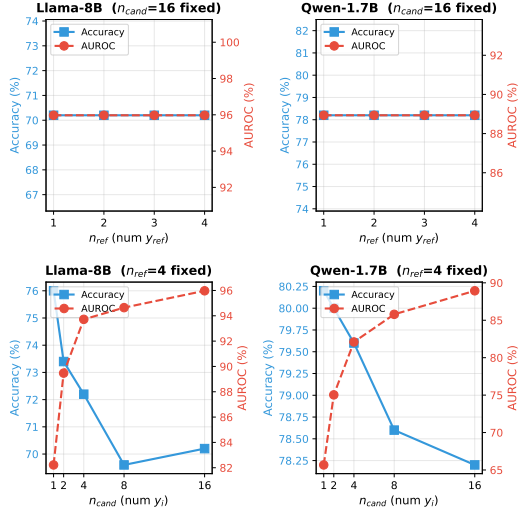


Figure 2: Effect of the numbers of pseudo-references (n_{ref}) and candidates (n_{cand}) on constrained selection for MorphMark on GSM8K.

WM	Method	Info $\uparrow$	Conc $\uparrow$	Faith $\uparrow$	Avg $\uparrow$
KGW	Baseline	3.55	4.84	4.38	4.26
	Q-only MBR	3.61	4.92	4.39	4.31
	Weighted	3.58	4.90	4.36	4.28
	Ours	3.62	4.91	4.45	4.33
SWEET	Baseline	3.70	4.93	4.31	4.31
	Q-only MBR	3.70	4.94	4.43	4.36
	Weighted	3.62	4.95	4.41	4.33
	Ours	3.63	4.97	4.36	4.32
MorphMark	Baseline	3.57	4.97	4.36	4.30
	Q-only MBR	3.56	4.95	4.39	4.30
	Weighted	3.68	4.90	4.36	4.31
	Ours	3.68	4.91	4.38	4.32
<i>Overall (averaged across all watermarks)</i>					
	Baseline	3.60	4.91	4.35	4.29
	Q-only MBR	3.62	4.94	4.40	4.32
	Weighted	3.62	4.91	4.38	4.31
	Ours	3.64	4.93	4.40	4.32

Table 4: Text quality evaluation on CNN/DM.

scale (5 is the best) for informativeness (Info.), conciseness (Conc.), and factual consistency (Faith.). Table 4 shows that our method remains competitive in text quality, tying for the highest average score overall while substantially improving watermark detectability in the main experiments.

D Performance on Robustness

Because lexical and semantic perturbations can weaken watermark signals, we evaluate attack robustness on CNN/DM under three attacks. **Word Del.** randomly deletes 10%, 20%, or 30% of words from the generated text, **Synonym** randomly replaces 10%, 20%, or 30% of words with WordNet-based synonyms, and **Paraphrase** rewrites each se-

WM	Model	Dataset	Baseline	Q-only MBR	Ours
KGW	Llama	CNN/DM	123.0	120.6	121.6
		XSum	44.3	41.7	42.5
		GovReport	224.1	229.9	230.5
		GSM8K	87.1	76.4	77.6
	Qwen	CNN/DM	137.9	135.4	136.9
		XSum	46.5	44.9	46.4
		GovReport	150.8	150.3	146.8
		GSM8K	140.0	138.3	141.9
SWEET	Llama	CNN/DM	129.9	125.1	126.8
		XSum	44.0	42.4	43.5
		GovReport	234.8	232.4	232.3
		GSM8K	90.9	76.5	85.3
	Qwen	CNN/DM	137.0	135.1	136.4
		XSum	46.9	44.8	45.3
		GovReport	152.7	151.0	154.4
		GSM8K	142.1	137.7	148.7
MorphMark	Llama	CNN/DM	123.3	121.8	122.8
		XSum	43.4	43.2	42.9
		GovReport	231.0	226.4	226.7
		GSM8K	78.6	78.0	78.4
	Qwen	CNN/DM	137.1	137.3	136.9
		XSum	46.7	47.8	47.1
		GovReport	149.9	154.3	151.1
		GSM8K	141.5	142.0	141.4

Table 5: Average output length per method.

lected output once with Phi-3.5-mini-instruct (Abdin et al., 2024) as a third-party paraphraser. Figure 3 shows that our method generally retains stronger detectability than baseline decoding and quality-only MBR across watermarking methods, models, and attack strengths, including the paraphrase attack.

E Output Length Analysis

Because the z-score in Eq. 1 depends on sequence length T , improved detectability could in principle arise from selecting longer outputs rather than more watermark-friendly ones. Table 5 reports average output lengths measured in whitespace-delimited tokens for each method. Across most settings, Ours remains close to Baseline and Q-only MBR and does not exhibit a systematic tendency to produce longer outputs. In many cases, it is slightly shorter than Q-only MBR. These results suggest that the detection gains of constrained selection are unlikely to be driven primarily by length inflation.

F Case Study

To illustrate how constrained selection operates at the document level, we examine a CNN/DailyMail example under KGW watermarking with Llama-3.1-8B. We use a candidate pool of size $K = 16$ and a keep ratio of $r = 0.7$, which retains the top 12 candidates according to their detection scores. The following examples show the outputs selected

Attack Robustness on CNN/DM (TPR@FPR ≤ 1%, %)

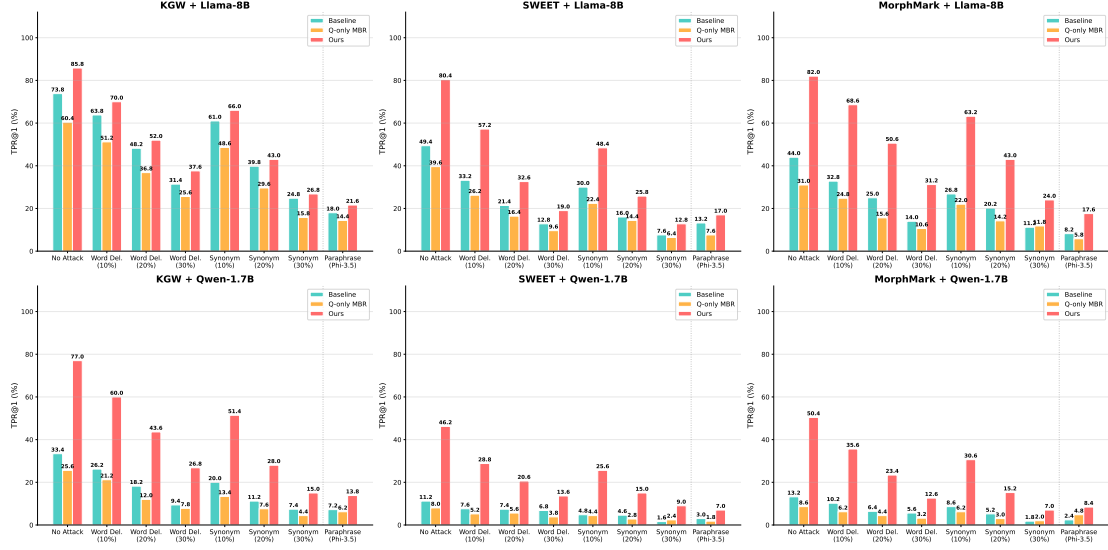


Figure 3: Attack robustness on CNN/DM (TPR at FPR ≤ 1%) under word deletion and synonym substitution attacks at 10%, 20%, and 30% corruption rates, and under a single paraphrasing attack.

by each strategy, together with their z-scores, green-token rates, and reference-based BERTScores.

While the baseline output exhibits a high watermark signal, Q-only MBR prioritizes semantic similarity and selects a candidate with a substantially lower detection score. In contrast, our method preserves a strong watermark signal while achieving comparable semantic quality, illustrating how constrained selection balances watermark detectability and generation quality.

G Quality-Detection Correlation

Figure 4 shows the per-candidate relationship between pseudo-reference quality (BERTScore against pseudo-references) and detection score (z-score) on CNN/DM with Llama-3.1-8B. Across all three watermarking methods, the two are negatively correlated (KGW: $r = -0.32$, SWEET: $r = -0.41$, MorphMark: $r = -0.23$; all $p < 10^{-90}$), showing that candidates with higher pseudo-reference quality tend to have lower detection scores. This pattern helps explain why quality-only reranking can weaken watermark signals in our setting.

H Why Constraints Can Improve over Scalarization

We provide a simple pairwise condition illustrating why constrained selection can be preferable to weighted scalarization when detectability is an acceptability requirement. For simplicity, let Q

and D denote the normalized quality and detection scores used in the weighted objective. Consider two candidates a and b that both satisfy the detection floor, i.e., $D(a) \geq \tau$ and $D(b) \geq \tau$. Suppose that $Q(a) > Q(b)$ but $D(b) > D(a)$. Under constrained selection, the additional detection score of b provides no advantage once both candidates are feasible, so the rule prefers a because it has higher quality. In contrast, weighted MBR prefers b over a whenever

$$(1-\lambda)Q(b) + \lambda D(b) > (1-\lambda)Q(a) + \lambda D(a), \quad (6)$$

or equivalently,

$$\lambda > \frac{Q(a) - Q(b)}{(Q(a) - Q(b)) + (D(b) - D(a))}. \quad (7)$$

Thus, even after the detection requirement has already been met, weighted scalarization can still sacrifice quality to further increase the detection score. Constrained selection avoids this behavior by using detection only as a feasibility condition and optimizing quality within the feasible set. This does not imply that constrained selection dominates weighted MBR in all settings. Rather, it identifies a setting in which the constrained objective is better aligned with the deployment goal: satisfy a minimum detection requirement first, and then maximize quality among acceptable candidates.

This condition is especially relevant when Q and D are negatively correlated over the candidate pool. In that regime, increasing λ tends to move selection globally away from high-quality candidates,

Case study on CNN/DailyMail

Green and red tokens indicate membership in the watermark lists.

Baseline ($y_0, z = 3.67$, green 67.3%,
 $BS_{\text{ref}} = 0.915$)

Real Madrid assistant coach Fernando Hierro believes that the club can win a second consecutive Champions League title , a feat that no team has achieved in recent history . He points to the current squad 's strength and confidence following last season 's Champions League win , and notes that the team has a winning mentality similar to the club 's successful late 90 s and early 200 0 s teams .

Q-only MBR ($y_{14}, z = 0.48$, green 52.3%,
 $BS_{\text{ref}} = 0.924$)

Real Madrid assistant coach Fernando Hierro believes his team has the ability to win the Champions League for a second consecutive season . He credits the squad with having the confidence and strength to succeed , comparing them to the successful teams of the late 1990 s and early 2000 s .

Ours ($r = 0.7, y_1, z = 2.74$, green 63.9%,
 $BS_{\text{ref}} = 0.924$)

Real Madrid assistant coach Fernando Hierro believes that the club can win a consecutive Champions League title , as it is embedded in their " DNA ." He notes that the current squad has the strength and confidence to repeat the successes of the late 1990 s and early 2000 s .

whereas the constrained rule only removes candidates below the required detection floor.

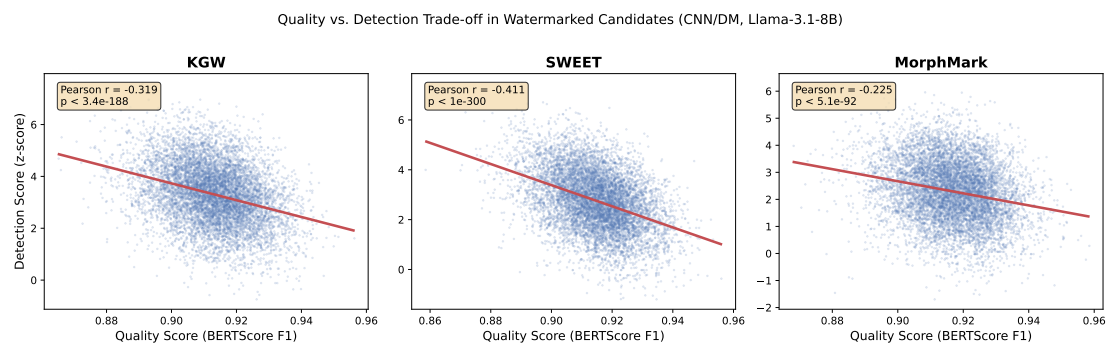


Figure 4: Per-candidate pseudo-reference quality versus detection score on CNN/DM with Llama-3.1-8B. Each point is one watermarked candidate.