WaterJudge: Quality-Detection Trade-off when Watermarking Large Language Models

Piotr Molenda, Adian Liusie, Mark J. F. Gales
ALTA Institute, Department of Engineering, University of Cambridge
pm725@cam.ac.uk, al826@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Watermarking generative-AI systems, such as LLMs, has gained considerable interest, driven by their enhanced capabilities across a wide range of tasks. Although current approaches have demonstrated that small, context-dependent shifts in the word distributions can be used to apply and detect watermarks, there has been little work in analyzing the impact that these perturbations have on the quality of generated texts. Balancing high detectability with minimal performance degradation is crucial in terms of selecting the appropriate watermarking setting; therefore this paper proposes a simple analysis framework where comparative assessment, a flexible NLG evaluation framework, is used to assess the quality degradation caused by a particular watermark setting. We demonstrate that our framework provides easy visualization of the quality-detection trade-off of watermark settings, enabling a simple solution to find an LLM watermark operating point that provides a well-balanced performance. This approach is applied to two different summarization systems and a translation system, enabling cross-model analysis for a task, and cross-task analysis.

1 Introduction

Large Language Models (LLMs) have progressed tremendously and are capable of generating high-quality texts for a diverse range of tasks. While these systems enhance automation, concerns arise about potential misuse, such as students using chat assistants for assignments or malicious users generating fake news articles. To counter this, current work has introduced the idea of LLM watermarking (Kirchenbauer et al., 2023a), where imperceptible patterns are injected into the generated text, enabling the statistical identification of whether text was generated by an LLM or not. However, most proposed watermarking schemes restrict the output generation space, which may lead to a trade-off

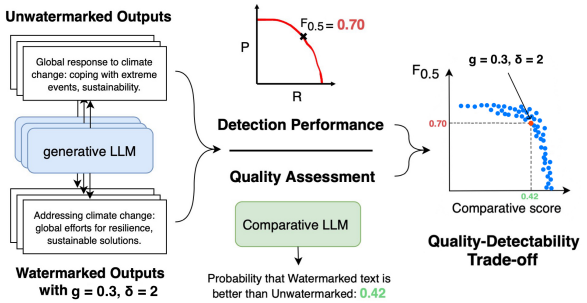


Figure 1: High-level overview of the WaterJudge Framework: Given a system, watermarking parameters, and set of inputs, watermarked outputs are assessed in terms of quality and detectability, leading to a curve over all operating points.

between quality and watermarking detection performance. Although there has been great effort into improving watermarking schemes for LLMs (Yoo et al., 2023; Kuditipudi et al., 2023; Kirchenbauer et al., 2023b), less work has analyzed the resulting quality degradation. It is common for watermarking schemes to measure quality by reporting the perplexity from a larger pre-trained LLM (Kirchenbauer et al., 2023a; Takezawa et al., 2023; Wang et al., 2023; Zhao et al., 2023; Ren et al., 2023; Liu et al., 2023), or to report similarity metrics such as BLEU or ROUGE(Fu et al., 2023; Takezawa et al., 2023; Li et al., 2023; Kirchenbauer et al., 2023b), however, these metrics are simplistic heuristics and may not truly capture actual output text quality, as discussed by Zhong et al. (2022); Wang et al. (2022); Zheng et al. (2023).

This work proposes WaterJudge, a framework for analyzing the trade-off between watermarking detectability and the quality of generated watermarked text. We leverage the LLM-as-a-judge evaluation approaches (Zheng et al., 2023; Liusie et al., 2023) to measure the average probability that an LLM prefers a watermarked text over an unwatermarked text. This is used as a metric for quantifying the quality degradation caused by watermarking, which with watermark detection performance,

can be used to determine the quality and detectability of a watermark operating point. This provides an approach for practitioners to visualize the effectiveness of specific watermarking operating points, enabling simple selection of an optimal watermark setting with minimal quality degradation.

2 WaterJudge

2.1 Soft-Watermarking Scheme

Language Models predict the conditional distribution of the next token $w_{i+1} \in \mathcal{V}$ given the input text $x_{1:M} \in \mathcal{V}^M$ and the previously generated tokens, $w_{1:i}$. For identification of LLM generated text, Kirchenbauer et al. (2023a) propose a simple soft-watermarking scheme, where the previous token w_i is used in a hash function to split the vocabulary into a mutually exclusive green list $\mathcal{V}_{\mathtt{g}}(w)$ and red list $\mathcal{V}_{\mathtt{r}}(w)$. The approach then incentivizes green-list words to be generated at the next step, such that the green-list word count can determine whether a text was generated by the LLM or not. The parameter g sets the relative size of the green list, such that $|\mathcal{V}_{\mathtt{g}}| = g \cdot |\mathcal{V}|$ and $|\mathcal{V}_{\mathtt{r}}| = (1-g) \cdot |\mathcal{V}|$. The watermarking scheme then increases the logits of all tokens in the green list by a bias δ,

$$l_k^{wm} = \begin{cases} l_k^{lm} + \delta, & \text{if } w_k^{lm} \in \mathcal{V}_{\mathtt{g}} \\ l_k^{lm}, & \text{otherwise} \end{cases} \quad (1)$$

Where l_k is the logit for the k'th token in the vocabulary w_k^{lm}. The watermarking scheme therefore has two parameters, the green list size s and the green list bias δ. The watermark score for a particular text is then calculated as the number of green list words present in the output text, where the higher the score, the more likely the output was generated by the watermarked LLM.

$$s_{\mathtt{wm}} = \frac{1}{N_w} \sum_{i=1}^{N_w} \mathbb{1}(w_i \in \mathcal{V}_{\mathtt{g}}(w_{i-1})) \quad (2)$$

Where $\mathbb{1}$ is the indicator function. Note that this watermarking scheme has the useful property that detection can be achieved even without model access. One only requires knowledge of the tokenizer and hashing function as this enables the green and red lists to be dynamically calculated, which is all that's needed to score texts. Further, if multiple models share a tokenizer, there could be an agreed watermarking convention that enables universal watermark detection over a range of models.

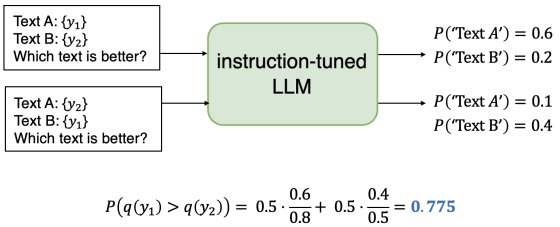


Figure 2: Comparative assessment probabilities are attained by calculating the likelihood of generating 'Text A' or 'Text B', normalizing, and averaging over both permutations. Example prompts are displayed, with the actual prompts shown in Appendix A.

2.2 Zero-shot Comparative Assessment

LLM comparative assessment (Liusie et al., 2023; Zheng et al., 2023), which prompts an LLM to determine which of two texts is better, is used in our framework to measure the quality degradation caused by watermarking. This method was selected due to being simple, zero-shot, and easily transferable to a range of tasks, as well as demonstrating impressive NLG evaluation performance.

For a given task and model, let x represent the input text, y the generated output text, and $y_{\mathtt{wm}}$ an output text generated from the system when watermarked. The Comparative assessment uses open-sourced instruction-tuned LLMs by querying which of the two provided texts is better. The comparative assessment system outputs $P(q(y_1) > q(y_2)|x)$, the probability that the quality of text y_1 is better than the text y_2, as demonstrated in Figure 2. The watermark degradation is measured over a corpus of input texts $\mathcal{D} = \{x^{(i)}\}_{i=1...N_d}$, with the average comparative selective probability used as the quality metric

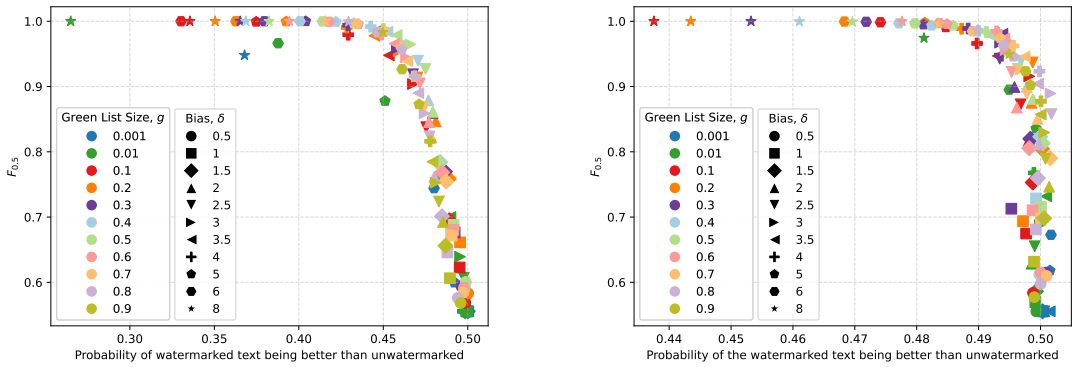
$$s_{\mathtt{q}} = \frac{1}{N_d} \sum_{i=1}^{N_d} P(q(y_{\mathtt{wm}}^{(i)}) > q(y^{(i)})|x^{(i)}) \quad (3)$$

where $y^{(i)}$ and $y_{\mathtt{wm}}^{(i)}$ are the generated base and watermarked outputs respectively, given input $x^{(i)}$.

3 Experimental set up

3.1 Datasets

We analyze the trade-off between quality and detection performance for two different tasks: summarization and translation. For the summarization task, 1024 contexts are sampled from the test set of XSumm (Narayan et al., 2018), while for translation 3072 German sentences are sampled from the test set of the XTREME corpus (Hu et al., 2020).



(a) BART

(b) Zephyr

Figure 3: The trade-off between quality and detectability when watermarking. Each point is a watermark setting with green list size g and bias δ, displaying $F_{0.5}$ detectability score and average Comparative Assessment probability.

3.2 Generative Models

Two different abstractive summarization systems are used; a BART-based summarization model trained on CNN-daily mail[1] (Lewis et al., 2020) and Zyphr-7B β instruction-tuned (Tunstall et al., 2023) which we prompt to perform summarization. For translation, mBART-large-50 is used, which is a BART model fine-tuned for multilingual translation[2] (Tang et al., 2020).

3.3 Watermarking Methodology

For each context, the model generates a baseline text without any watermark, and then multiple watermarked texts using various operating points. The watermarking operating points are taken by considering all combinations of green list size g ranging from 0.001 to 0.9 and bias δ ranging from 0.5 to 8. For summarization, the watermark score is the count of the fraction of green list words in the generated text, while in translation the output texts are grouped in sets of three (to achieve similar expected lengths for detection) and the score is computed for the grouped set. For each operating point, a threshold is chosen to classify watermarked and unwatermarked texts for the maximum $F_{0.5}$ value, which is then used as a detectability metric. $F_{0.5}$ is a weighted harmonic mean of precision and recall, giving more importance to precision than F_1 to safeguard against false positives. We use the same hashing seed to generate all green-lists, however, Appendix D shows that consistent results can be observed across random seeds.

[1] https://huggingface.co/facebook/bart-large-cnn

[2] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

3.4 Comparative Assessment Set Up

FLAN-T5 3B (Chung et al., 2022) is used as the base evaluation LLM, chosen due to its demonstrated pairwise evaluation abilities (Liusie et al., 2023) and good multi-lingual capabilities. As the maximum length of the model is 1024 tokens, if the input prompt exceeds this limit, the end of the context is truncated to fit into the maximum limit, which avoids any of the summaries/translations being truncated. The comparative quality score is taken as the average of all $1024/3072$ samples.

4 Results

Summarization Figures 3a and 3b illustrate the relationship between summary quality and watermark detection performance for BART and Zephyr respectively. A clear trade-off between watermark strength and output quality can be observed for both systems, where strong watermarking degrades quality while weak watermarking maintains quality but yields poor detection performance. The results further suggest that though multiple operating points can yield similar quality-detectability characteristics, the framework provides a simple way to visualize points that achieve a good balance between the two. This can be useful for hyper-parameter selection, e.g. to find the setting where there's minimal quality degradation for a desired $F_{0.5}$ detectability score. Note that the quality scores are upper-bounded near 0.5, consistent with the idea that weak watermarking will enforce little restriction and yield texts of similar quality, while stronger watermarks will restrict generation and therefore yield texts of worse quality. Further, the saturation at $F_{0.5} = 1$ denotes the region where one can perfectly differentiate watermarked texts from unwatermarked texts, albeit often at the cost

of large quality degradation.

Additionally, it is observed that different base models can have varying optimal watermarking parameters. Zephyr-7B is much larger than BART (7.2B vs 0.4B parameters) and is likely to have a more accurate underlying task language model. As such, it seems to better deal with the restrictions imposed by watermarking, as seen by the vertical region around the probability of 0.5 (where minimal quality degradation and good detectability are achieved). Further, in the most extreme settings, Zephyr's average comparative probability drops to 0.44 compared to BART's 0.28. Examples of the generated watermarked text can be seen in Appendix H.

Translation We repeat analysis for translation, with Figure 4 showing similar quality-detectability characteristics when an mBART system, which translates German sentences to English, is watermarked. The plot shows further evidence of how for weaker models (0.6B parameters supporting 50 languages) strong watermarking can cause a significant drop where the system struggles to maintain quality. Additionally, we can observe that mBART is more sensitive to watermarking parameter settings and that quality is better for small-green list sizes than for larger-green list sizes, even for settings with equivalent detectability performance.

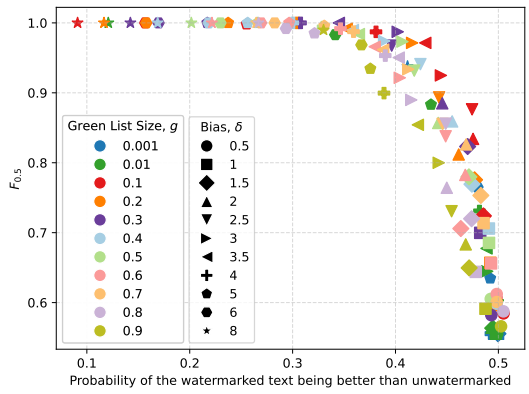


Figure 4: Results for watermarked translations generated with mBART for combinations of green list size g and bias δ.

Suitability of Comparative Assessment To verify that comparative assessment provides meaningful quality evaluation, we compare the generated quality scores against those from UniEval (Zhong et al., 2022) and COMET(Rei et al., 2020) which both demonstrate strong alignment with human judgment. UniEval is a summary assessment method using a T5-based boolean-answering system trained specifically to assess summaries on coherence, consistency, fluency, and relevance, while COMET is an open-source neural framework for machine translation evaluation. For summarization, comparative assessment has a Spearman correlation of **0.986** relative to UniEval scores[3], while in translation comparative assessment has a Spearman correlation of **0.988** relative to COMET. Figure 5 illustrates the relationship between the two quality scores of watermarked summaries generated by BART, with a similar graph for mBART translation shown in Appendix G). These results highlight that despite being simple and zero-shot, quality assessment via comparative assessment correlates highly with alternative high-performing automatic evaluation approaches that have been tailored to particular tasks. WaterJudge is a clear improvement over more dated metrics such as ROUGE or BLEU, which when used fail to capture the quality-detection trade-off (shown in Appendix C). Further, current popular methods such as perplexity have weaker correlations with UniEval and COMET (**0.922** for summarization and **0.940** for translation) and are more difficult to compare between models, where WaterJudge also shows additional promising capabilities, discussed in the next section.

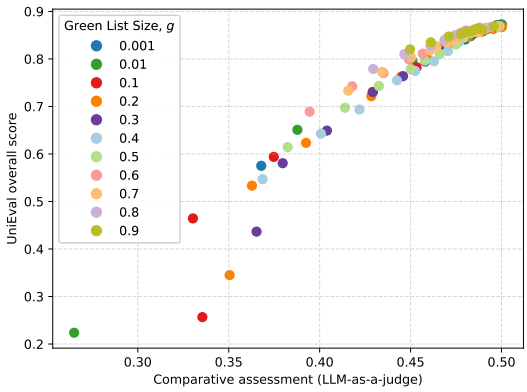


Figure 5: Scatter plot showing correlation between Comparative Assessment and UniEval for BART.

Transferability of Settings As an extension to the current analysis, we consider whether one can avoid doing a full grid search over all watermarking settings and instead transfer settings across different models and tasks. Firstly, it's observed that by looking at the expected quality scores of generated summaries for different operating points on BART and mBART, we observe a Pearson corre-

[3]scores of the 4 attributes are averaged as an overall score

lation $\rho = 0.927$ (Figure 6), while for BART and Zephyr, the correlation is $\rho = 0.826$. The lower correlations for BART-Zephyr can be explained due to the observed truncated linear relationship, where for weak watermarks Zephyr can apply watermarks without causing any quality loss, while for medium watermarks (e.g. $g = 0.5$) there remains a linear degradation to both systems (shown in Figure 14 in the Appendix). Using perplexity quality scores does not demonstrate strong cross-system correlations, and as shown in Figure 7, does not demonstrate the linear relationships that are observed with comparative assessment. Therefore perplexity scores may not be effective when considering transferring watermarking performance.

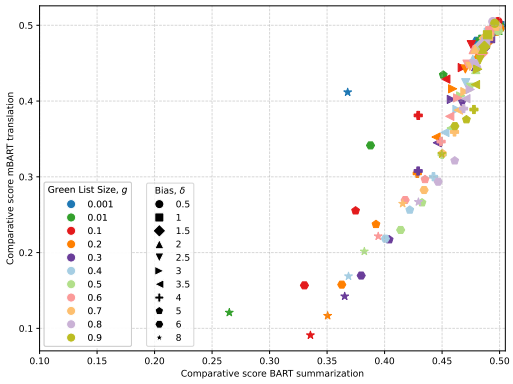


Figure 6: Relationship of watermark settings' comparative assessment quality scores for BART and mBART.

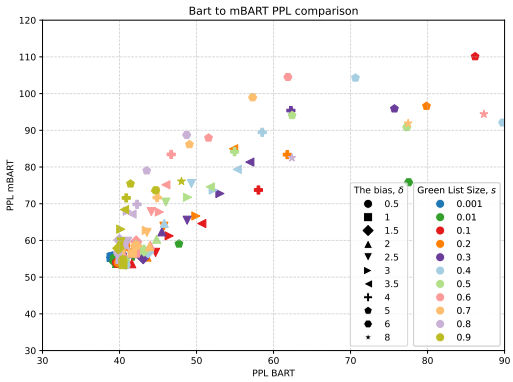


Figure 7: Relationship of watermark settings' perplexity scores for BART and mBART (ignoring the outlier, very high perplexity, points).

Moreover, the $F_{0.5}$ scores of watermark settings on different models are also highly correlated: BART-Zephyr quality scores have PCC $\rho = 0.986$, while for BART-mBART the PCC is $\rho = 0.990$. Even though this is a cross-task comparison, BART and mBART have a near 1:1 mapping in detectability scores (Figure 8) while Zephyr detectability scores tend to be slightly higher than those from BART (which is mostly due to length mismatches, as discussed in Appendix B). The high linear correlations for both quality and detection suggest that WaterJudge can be used to map performance on one model/task to another, which may yield additional predictive abilities for generating the full detectability-quality trade-off curves. Initial examples of the effectiveness of transferring settings across systems are shown in Appendix F.

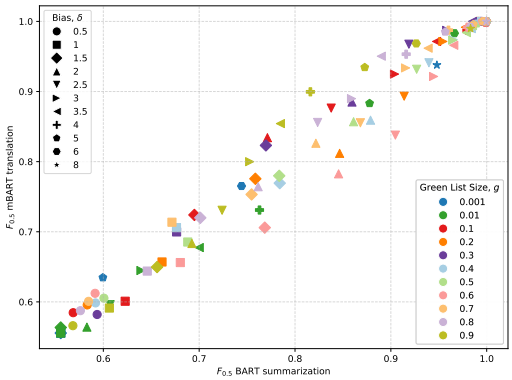


Figure 8: Comparison of watermark settings' detectability scores for BART and mBART.

5 Conclusions

This paper introduces WaterJudge, a framework for investigating the quality-detection trade-off when watermarking LLMs, enabling easy visualization of various watermarking settings and simple hyperparameter selection. Comparative Assessment is shown to be a practical metric for measuring quality degradation and improves on currently used evaluation methods in its accuracy and versatility. WaterJudge is also useful in cross-task and cross-model analysis, showing good correlations for both detectability and quality, despite varying characteristics due to model strength.

6 Limitations

Although LLM evaluation approaches have recently been demonstrated to be effective reference-free evaluation methods, there may be inherent biases such as self-enhancement bias that can impact the robustness of the approach and cause discrepancies in human evaluation. This study could further investigate sensitivity to evaluation prompt sensitivity, or output length, as well as extend to more models, watermarking schemes, and tasks.

7 Ethical Concerns

Watermark detection performance may not be completely accurate, and false negatives may lead to individuals being unfairly charged for using AIs, when they may have written the text themselves.

8 Acknowledgements

This work is supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023a. 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.

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

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yuhang Li, Yihan Wang, Zhouxing Shi, and Cho-Jui Hsieh. 2023. Improving the generation quality of watermarked large language models via word importance scoring. *arXiv preprint arXiv:2311.09668*.

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

Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2023. Zero-shot nlg evaluation through pairware comparisons with llms. *arXiv preprint arXiv:2307.07889*.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2023. A robust semantics-based watermark for large language model against paraphrasing. *arXiv preprint arXiv:2311.08721*.

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

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Towards codable text watermarking for large language models. *arXiv preprint arXiv:2307.15992*.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2022. Perplexity from plm is unreliable for evaluating text quality. *arXiv preprint arXiv:2210.05892*.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust multi-bit natural language watermarking through invariant features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2115.

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

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

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Prompts

Use case	Prompt
Zephyr summarization	<|system|> You are a tool providing a short text summary. <|user|> Write a short summary of the following text: context <|assistant|>
FLAN-T5 summarization Comparative Assessment	Passage: {passage} Summary A: {summary 1} Summary B: {summary 2} Between Summary A and Summary B, which text summarises the passage better?
FLAN-T5 translation Comparative Assessment	Original text: {context} Translation A: {translation 1} Translation B: {translation 2} Between Translation A and Translation B, which is the better translation of original text?

Table 1: prompts used for experiments.

For reproducibility, Table 1 shows the prompts used for summary generation (using Zephyr 7B β) and comparative assessment (with FLAN-T5 as the base LLM). For summarization, we evaluated the overall summary quality, as in initial experiments where particular attributes were assessed, the LLM struggled to differentiate between the different attributes with simple prompts (e.g. 'fluency', 'coherence', 'consistency', or 'relevance'). We use a Tesla V100S 32Gb GPU to conduct all experiments. FLAN-T5 Comparative assessment takes 6 minutes to assess each summarization watermark operating point (1024 samples) and 10 minutes to assess each translation operating point (3072 samples). It takes 5 minutes for BART to generate 1024 summaries, 40 minutes for Zephyr-7B β to generate 1024 summaries, and MBART 12 minutes to generate 3072 translations.

B Watermarked texts length

Figures 9, 10 and 11 show the average lengths (in tokens) of the outputs of the models. BART and mBART were fine-tuned for a specific task and therefore the outputs typically have consistent length (usually 60-80 tokens). Zephyr 7B β tends to generate longer summaries with a larger variance in the output lengths. Note that longer texts will typically be easier to detect since having more generated words will reduce the expected variance from the expected fraction of green list words. Therefore when choosing optimal operating points, one should also take the length into account.

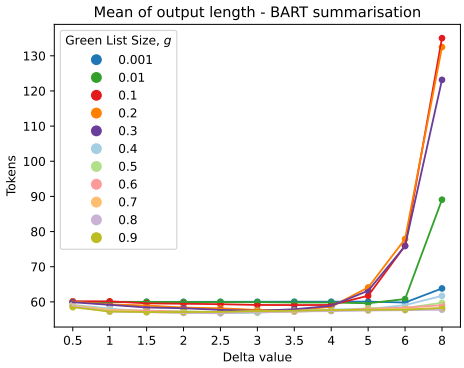


Figure 9: Average length (in tokens) of output BART (summarization) texts for various watermark settings.

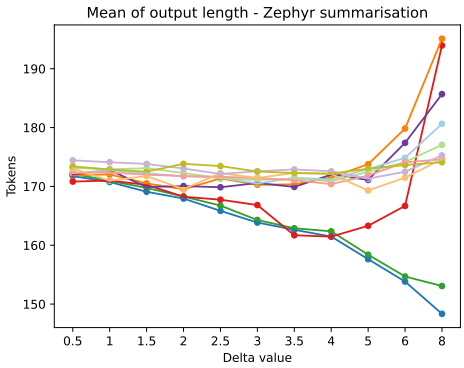


Figure 10: Average length (in tokens) of output Zephyr (summarization) texts for various watermark settings.

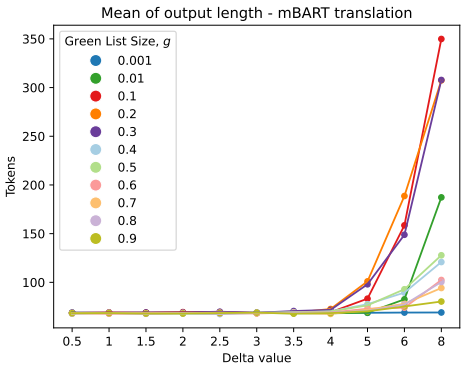
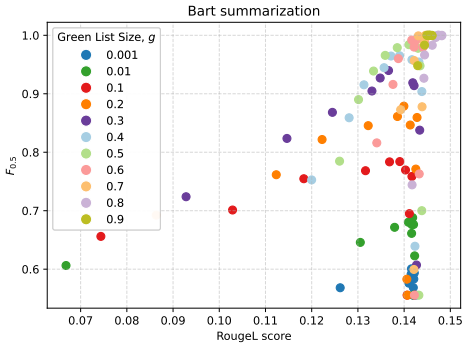


Figure 11: Average length (in tokens) of output mBART (translation) texts for various watermark settings.

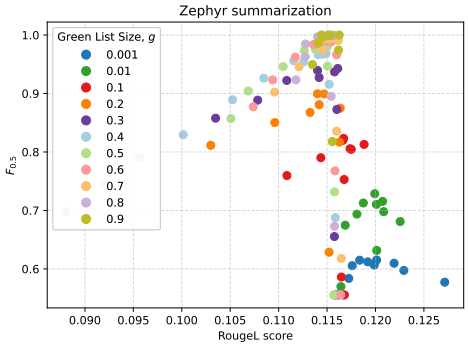
Moreover, the average lengths for models show an issue with simple watermark partitioning into green/red lists: if '</s>' is in the red list of '.', then the outputs tend to grow overly long (since ending the sequence incurs a red-list word). In our experiments, green lists are subsets of larger green lists, and, for both models, the aforementioned issue occurs when $g < 0.4$. High bias δ texts with $g < 0.4$ are significantly longer than those from other settings, which considerably impacts text quality (though for very small green lists, $g \leq 0.001$, there are more red-list words generated and therefore the sentence may end as expected). For the given randomized seed, Zephyr does not have eos token '</s>' in green list of '.' (for any $g > 0.9$). Hence, the rise is visible for all larger green lists $g \geq 0.1$, but the issue is not as significant as for the other models (due to its better capability to adapt to watermark restrictions). This highlights that this problem is significant when evaluating very strong bias operating points (with generally unusable outputs), but does not otherwise influence evaluation (see Appendix D).

C Baseline Evaluation Metrics

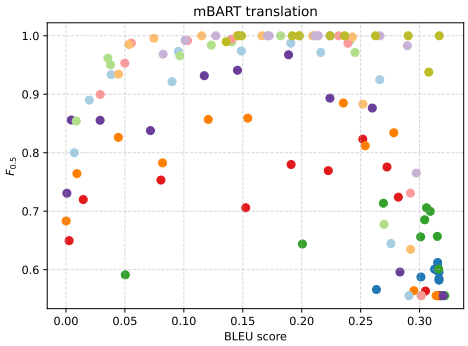
Instead of using comparative assessment to assess the quality degradation, we generate equivalent plots using metrics such as ROUGE or BLEU (against reference summaries/translations), which are standard watermark evaluation metrics. Figure 12 shows that using these evaluation metrics leads to curves that mask the quality-detection tradeoff and provide little insight. The RougeL curves seem to be strongly influenced by summary length (see Figures 9, 10, 11), while the BLEU metric has little explanation. This highlights that the WaterJudge framework requires a capable and effective evaluation approach to capture the quality-detection trade-off and that comparative assessment is a suitable method.



(a) $F_{0.5}$ against RougeL scores for BART summarization.



(b) $F_{0.5}$ against RougeL scores for Zephyr summarization.



(c) $F_{0.5}$ against BLEU scores for mBART translation.

Figure 12: Quality-Detectability trade-off curves for commonly used similarity metrics.

D Green List Seed Consistency

Results in Appendix B suggested that there may be some seed variability, dependent on specific word (or special token) green list bi-grams. To verify the consistency of our results, evaluation for three different seeds is shown in Table 2. The seeds were selected to maximize variability, such that for seed 1 '</s>' is never in the green list of '.',

for seed 2 it occurs when $g > 0.4$, and for seed 3 '</s>' is always in the green list of '.' It's observed that even in these settings, there is little impact on the metrics for the main operating points (δ is 3 or 6), and only when in regions with very heavy watermarks ($\delta = 9$) are small differences seen. It is worth noting that the length is affected by the seeds, but it's not necessarily negatively received by the Comparative Assessment (in contrast to metrics like Rouge). Due to this designed length bias, seed 1 does on average report slightly higher $F_{0.5}$.

		$F_{0.5}$			Quality		
g	δ	1	2	3	1	2	3
0.2	3	0.90	0.86	0.88	0.44	0.44	0.44
0.5	3	0.90	0.89	0.86	0.44	0.44	0.43
0.8	3	0.76	0.73	0.72	0.46	0.45	0.45
0.1	6	1.00	1.0	1.0	0.28	0.29	0.29
0.4	6	1.00	1.0	1.0	0.35	0.34	0.33
0.7	6	0.99	0.99	0.99	0.39	0.40	0.39
0.01	9	1.0	1.0	1.0	0.15	0.12	0.12
0.2	9	1.0	1.0	1.0	0.16	0.15	0.16
0.3	9	1.0	1.0	1.0	0.23	0.19	0.20
0.6	9	1.0	1.0	1.0	0.31	0.30	0.28

Table 2: Table comparing detectability and quality scores of three additional seeds for various operating points in BART summarization , with good agreement between all seeds.

E Model to model comparison

Figures 13 details the relationship of *detectability* for different watermarking settings for BART-Zephyr, while Figure 14 shows the equivalent graphs for *quality*. Figure 15 shows the BART to Zephyr comparison for Perplexity, where notably most of the points are tightly grouped in a single region. Note that highly hallucinated outputs have extremely high perplexity scores, and so have been cropped out of the plot. Moreover, Figure 14 suggests that for most models, large green list sizes can yield reasonable detectability with minimal quality degradation, medium green list sizes have predictable and transferable linear degradation, while small/very small green lists have unpredictable behavior and should be avoided.

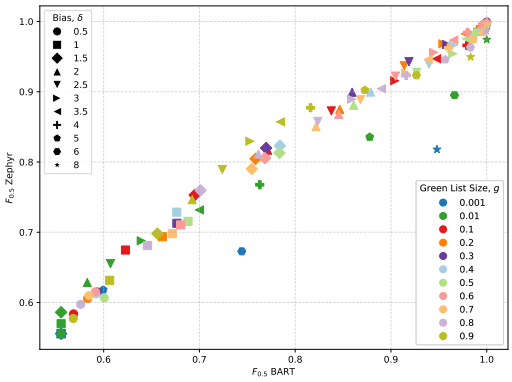


Figure 13: Comparison of watermark settings' detectability scores for BART and Zephyr

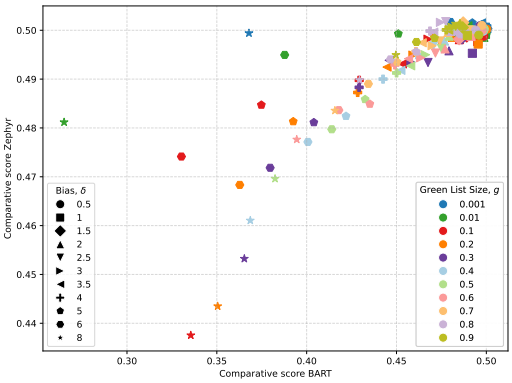


Figure 14: Comparison of watermark settings' quality scores for BART and Zephyr

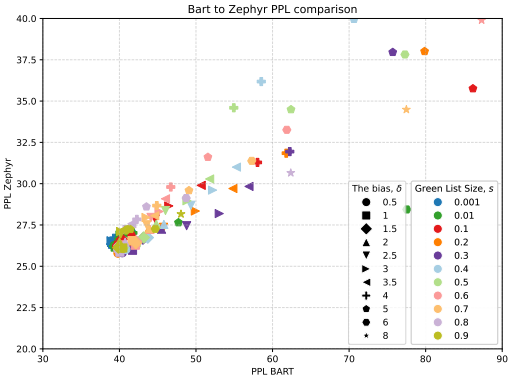


Figure 15: Comparison of watermark settings' PPL for BART and Zephyr (ignoring the outlier, very high perplexity, points).

F WaterJudge predicitive capabilities

Appendix E and Section 4 demonstrated that both Zephyr-BART and mBART-BART have consistent and linear relationships between quality and predictive scores. Detectability is nearly equivalent across different watermarking settings, starting from $F_{0.5} = 0.5$ and linearly increasing to

$F_{0.5} = 1$. Alternately, the Quality comparisons can be broken into three regions: for weak watermark settings, large models (like Zephyr) can maintain quality, for medium strength watermarks (e.g. $g = 0.5$) there is a linear degradation for all systems, while strong watermarks can lead to meaningless output texts and low transferability (deviations from the trend). Meaningful regions of both of these curves can be estimated well with a two-parameter function (such as truncated at the top linear function), which enables a transformation of watermark performance from one system to another while using only a few tested operating points. To achieve the fitting, a parameterized hyperbolic tangent was fitted to the BART quality-detection curve, as shown in Figure 16, by minimizing the average perpendicular Mahalanobis distance of the operating points from the curve. This has been done to get a smooth baseline function capturing the whitened data shape.

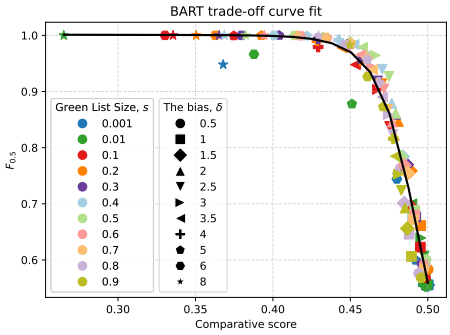
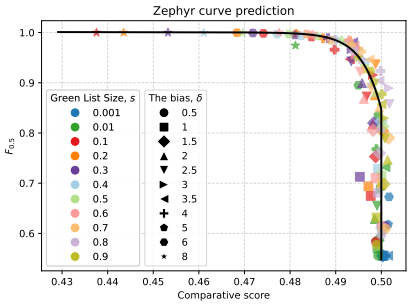


Figure 16: Curve fit to BART operating points graph.

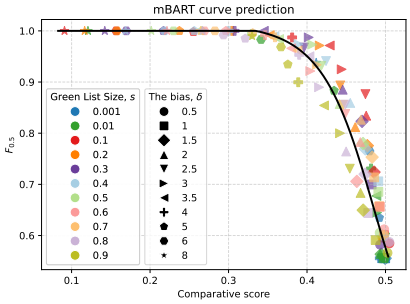
By fitting truncated linear functions to the relationships such as Figures 13-14, one can transform a baseline curve to achieve predicted fits, as shown in Figure 17. These 'predicted' shapes are achieved with significantly fewer points and avoid the grid search, which can be useful when attempting to determine whether an effective watermark setting exists for the new system, and also enable fast and thorough testing across hyperparameters and models.

G COMET Comparison

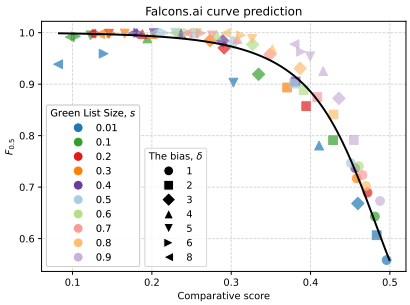
Figure 18 shows how Comparative Assessment for translation is also strongly correlated with the most recent automatic translation evaluation metric: COMET. The Spearman correlation of quality scores is **0.988**.



(a) Zephyr



(b) mBART



(c) Falcons.ai

Figure 17: Estimated LLM Quality-Detectability trade-off curves from model-to-model comparisons.

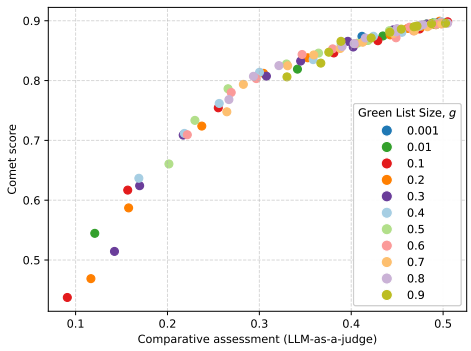


Figure 18: Correlation between Comparative Assessment and COMET score for mBART translation.

H Examples of watermarked texts

Table H provides examples of possible output texts under various operating points for the two summarization experiments for a sampled article. The texts achieve various detectability and show varied degradation in text quality.

Context	The 41-year-old reached the semi-final for the first time since he last lifted the trophy in 2011 with a 13-6 win over Kyren Wilson. The world number six will now face the winner of Stephen Maguire's match against Barry Hawkins for a place in Sunday's final. "I believe I can [win], and that's a great feeling to have," Higgins said. "I was close last year, Alan [McManus] pipped me [in the quarter-final]," he added. "A couple of years after 2011, I'd been losing, my form wasn't great and you don't think you're [continues]
Unwatermarked BART	Higgins beat Kyren Wilson 13-6 to reach the semi-final of the World Championship. The Scot will face the winner of Stephen Maguire's match against Barry Hawkins. The 41-year-old believes he can win the tournament for the first time since 2011.
$g = 0.7, \delta = 6$ $F_{0.5} = 0.996$ Quality score 0.434	Higgins beat Kyren Wilson 13-6 in the quarter-finals at the World Cup. He now awaits the winner of Stephen Maguire's match against Barry Hawkins. It is the first time the 41-year-old Scot is in the semi-final since he lifted the trophy in 2011.
$g = 0.5, \delta = 4$ $F_{0.5} = 0.988$ Quality score 0.450	Scottishman Higgins is into the semi-final for the first time since 2011. He beat Kyren Wilson 13-6 in his quarter-final on Wednesday. He now faces the winner of Stephen Maguire and Barry Hawkins. He says he has more self-confidence heading to final.
$g = 0.1, \delta = 8$ $F_{0.5} = 1.00$ Quality score 0.336	Bobbyiggins into first semifinal since picking off David Higgins 11 years ago. 41-year-old Scot defeated World 16's KYRN WEIRD 12 4 8 3 3 3 8 1 this night at BAVCA PIA Stadium. Beats Ian Pinn are first-round bye; winner of Stephen MagUires takes part Saturday afternoon and it goes 6 vs 6 - winner to meet for the tournament [continues]
Unwatermarked Zephyr	John Higgins, aged 41, advanced to the snooker World Championship semi-finals for the first time in a decade with a 13-6 victory over Kyren Wilson. Higgins will face either Stephen Maguire or Barry Hawkins for a place in the final, and believes he can win the tournament. The Scot has won all three sessions against Wilson and [continues]
$g = 0.5, \delta = 4$ $F_{0.5} = 0.984$ Quality score 0.491	John Higgins, 41, progressed to his seventh World Snooker Championship semi-final, beating Kyren Wilson 13-6. Higgins, ranked sixth, aims for his fourth title after last lifting the trophy in 2011. The Scottish player faces either Stephen Maguire or Barry Hawkins in the last four, with Higgins preferring a Maguire match as a fellow [continues]
$g = 0.1, \delta = 8$ $F_{0.5} = 1.00$ Quality score 0.438	John Higgins, aged 41, advanced to the world snooker championship semi-finals for the first time since lifting the trophy in 2011. In the quarter-finals, Higgins defeated Kyren Wilson in three sessions by a scoreline of 13-6. In the semi-finals, Higgins will confront either Stephen Maguire or Barry Hawkins for a position in the final on Sunday. Reflecting after his match against Wilson, Higgins stated that he "believ(ed) (he) [continues]

Table 3: Examples of watermarked and unwatermarked outputs for a few chosen operating points for summarization. It may be noted that in the strong watermarking region, the model begins 'hallucinating' or repeating text continuously. This yields long, unusable texts with high density of green list words, making the operating points a bit less meaningful for analysis. This is more difficult to trigger for larger models like Zephyr.