

Appendix A. Bounding Output Differences with LDP

Let \mathcal{A} be a randomized mechanism, R_θ be the model parameter space, and ϵ denote the privacy budget associated with the noise scale δ . In δ -STEAL, we add LDP noise $Lap(0, \delta)$ to the token embeddings, which is equivalent to applying ϵ -LDP to each data sample. By the post-processing property of LDP, the adversary’s model θ_{adv} is also ϵ -LDP. As a result, watermark detectors cannot determine whether θ_{adv} was trained on watermarked data and fail to verify its intellectual property, especially with small values of ϵ . The difference between θ_{adv} trained with or without watermarked data is bounded as follows: $P[\mathcal{A}(x + Lap(0, \delta), y^{wm} + Lap(0, \delta)) \in R_\theta] \leq \exp^\epsilon P[\mathcal{A}(x' + Lap(0, \delta), y'^{wm} + Lap(0, \delta)) \in R_\theta]$.

Appendix B. Privacy Budget Calculation

In δ -STEAL, we use a common LDP approach, which is a Laplace mechanism that adds Laplace noises into token embeddings of the model. The Laplace mechanism is defined as: $\mathcal{A}_\mathcal{E}(x, \mathcal{E}(x), \epsilon) = \mathcal{E}(x) + (L_1, L_2, \dots, L_d)$, where $\mathcal{E}(x)$ is the embedding of token x , d is the embedding size, and L_i are independent and identically distributed (i.i.d.) random variables drawn from Laplace noise centered at 0, scaled by $\delta = \frac{\Delta(\mathcal{E})}{\epsilon}$. Given a noise scale δ , to compute the privacy budget ϵ , we first need to calculate the magnitude by which a single individual’s data can change the function \mathcal{E} in the worst case, which is defined as: $\Delta(\mathcal{E}) = \max_{\forall x, \tilde{x} \in N^d} \|\mathcal{E}(x) - \mathcal{E}(\tilde{x})\|_1$.

For LLaMA-2, we obtain $\Delta(\mathcal{E}) = 0.3$, while for Mistral, $\Delta(\mathcal{E}) = 0.05$. With noise scales of $\delta \in 0.001, 0.01, 0.05, 0.1$, the corresponding privacy budgets are:

- For LLaMA-2: $\epsilon = \frac{\Delta(\mathcal{E})}{\delta} = \{300, 30, 6, 3\}$
- For Mistral: $\epsilon = \frac{\Delta(\mathcal{E})}{\delta} = \{50, 5, 1, 0.5\}$

Appendix C. Supplemental Results

In this supplement, we include the results for the SemStamp watermark, which were not presented in the main body. SemStamp requires fine-tuning a robust sentence embedder to support the local sensitivity mechanism for sentence selection. Since we used the pre-trained embedder provided by the authors, the results are not optimal in our experiments.

δ -STEAL against Semstamp. For Semstamp, our observations of δ -STEAL attacks in Fig. 6 are consistent with other watermarks. As noise scales increase, AttackSR rise, and PPL increases but remain comparable to the Baseline. However, a notable concern is the low Baseline watermark detection rate, which results in unexpectedly high AttackSR even without attacks, reaching 52.97% for LLaMA-2 and 59.15% for Mistral.

δ -STEAL and Existing Attacks on Semstamp. Fig. 7 illustrates the effectiveness of our δ -STEAL attacks compared to other attacks on Semstamp. It is clear that δ -STEAL, regardless of noise scale, performs effectively, indicating high AttackSR and low PPLs.

δ -STEAL in MMLU Downstream Task. In Table 3, we present only the KGW and EXP watermarking techniques, although we also experimented with all four techniques considered in this paper, including SIR and Semstamp. For SIR, our attempt to apply the publicly available trained model to the MMLU task resulted in poor accuracy, only 27.90% for LLaMA-2 and 20.90% for Mistral without any attacks. This poor performance is due to the fact that, due to computational complexity,

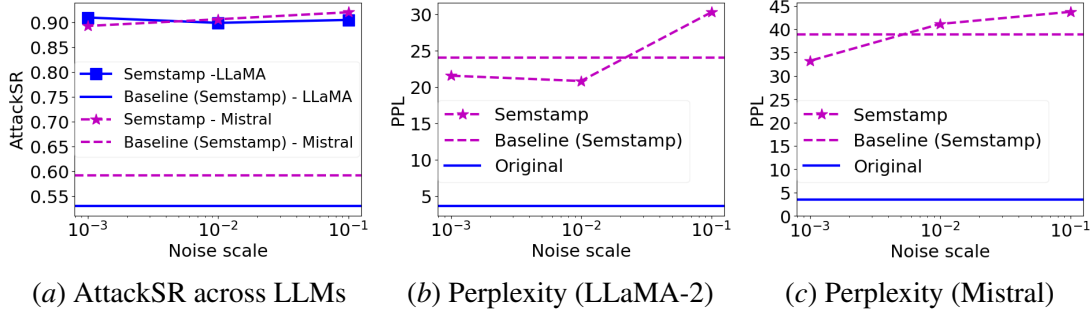
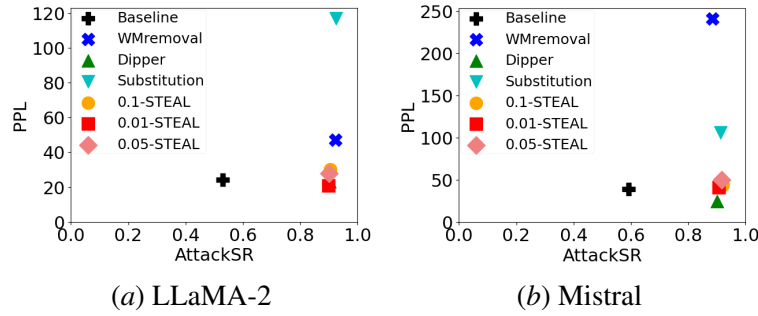
Figure 6: AttackSR and Perplexity of δ -STEAL on Semstamp.

Figure 7: AttackSR and Perplexity across on Semstamp.

we do not retrain their watermark model, meaning it may not adapt well to our data and settings. For Semstamp, which generates entire sentences instead of individual tokens, it is not suitable for MMLU, which involves multiple-choice question-answering tasks.

Semantic Preservation of δ -STEAL Outputs. We present additional quantitative examples from Mistral, as shown in Table 4a), to illustrate the effects of our δ -STEAL attack. These examples highlight the subtle differences introduced while preserving the overall semantic meaning. For instance, in the first row, the attack changes “office has talked to Attorney General Jeff Sessions” to “team has talked with Jeff Sessions” and removes the adverb “now,” while maintaining the key information about the Russia investigation. In the second example, “20 percent of black students and 10 percent of Latino students in Boston are attending the city’s top 20 public schools” is modified to “who applied to the city’s top schools were admitted,” subtly shifting the focus but retaining the core message that the mentioned students were accepted into top schools. Lastly, in the third instance, δ -STEAL changes “shape of a new toy” to “shape, configuration, and appearance of a new product” and modifies the quantifiers from plural to singular, changing “patent” and “variety,” while still discussing the types of patents. These examples demonstrate that our attack makes slight wording changes without altering the overall semantic content.

Table 4a): Examples of prompts and watermarked outputs with and without δ -STEAL (Mistral model). Green text highlights similarities and red text emphasizes differences.

Prompt	Watermarked (W)	Watermarked under δ -Steal (SA) ($\delta = 0.01$)	PPL (W)	PPL (SA)
... There have been a number of revelations this week related to the Russia investigation. Among them, that the special counsel's ...	KGW: office has talked to attorney general Jeff Sessions and that Robert Muller is now looking at sitting down with President Donald Trump. And, of course, news about a deal to temporarily end the government shutdown, but with no resolution [...continues]	KGW: team has talked with Attorney General Jeff Sessions, and that Robert Mueller is looking to sit down for an interview with President Trump. But with a government shutdown still hanging over everyone's heads, it can be easy to lose track of [...continues]	8.52	8.21
... Latino students as well saw less opportunity for access to quality schools than their white and Asian counterparts," O'Brien said. The study found that more than 80-...	EXP: percent of black students and 70-percent of Latino students in Boston are attending schools that are considered low performing. The report also found that 20-percent of black students and 10-percent of Latino students in Boston are attending the city's top 20 public schools. " The report [...continues]	EXP: percent of black students and 70-percent of Latino students who applied to the city's top schools were not admitted. The report also found that 20-percent of black students and 10-percent of Latino students who applied to the city's top schools were admitted. The report [...continues]	2.87	2.03
...patent protects the functional aspects of an invention, such as a new machine. A design patent protects the ornamental appearance of an invention, such as the...	SIR: shape of a new toy. Plant patents protect, you guessed it, new varieties of plants. You apply for a patent with the United States Patent and trademark Office, but you can not put a patent symbol "@" on your product until it is [...continues]	SIR: shape, configuration and appearance of a new product. A plant patent protects inventions of natural organisms such as a new variety of fruit tree. Once you determine what type of intellectual property you need, you can file a patent application with the U.S patent [...continues]	5.67	5.52

Table 4b) provides a comparison of our δ -STEAL with the different attack methods used in this study. While the approach of our δ -STEAL attack has been detailed previously, Dipper and the Substitution attacks present alternative solutions. As per Dipper, since it works with sentence levels, it truncates and discards all incomplete sentences within the target output to attack and always starts paraphrasing full sentences only, resulting in a disrupted reading flow from input to output and causing small loss of information. For example, in the first row, while both δ -STEAL and Substitution attacks continue to generate "JB" to complete "BJP", Dipper starts with a new sentence. Furthermore, due to Dipper's reliance on reordering sentences, the provided snippet of its output within a limited token count shown in Table 4b could not fully show relevant benchmark against other attacks. For instance, in the second example, the sentence "The company announced that it would pay \$50,000 for the silence of this alleged affair." introduces content beyond what is covered in the benchmark snippet. In contrast, the Substitution attack replaces texts by considering surrounding context, maintaining a high degree of similarity (as shown by the larger green portion) with the original watermarked text.

Table 4b): Examples of prompts and watermarked outputs of δ -STEAL and other attacks (LLaMA-2 model). Green text highlights similarities and red text emphasizes differences.

Prompt	Watermarked (W)	Watermarked under δ -Steal (SA) ($\delta = 0.01$)	Dipper	Substitution
... do that for the MCD polls, " said a senior party leader. Another reason behind not targeting Modi in the MCD polls is the massive victory the B ...	KGW: JP achieved in the 2014 Lok Sabha polls in Delhi , where the party won all seven seats. "We have to take into account the ground reality and the political equation in the city. We can not ignore the BJP 's victory in the 2014 Lok Sabha [...continues]	KGW: JP registered in the 2014 Lok Sabha elections in the city. " We have to convince the people of Delhi that why we are the best option to govern the city. We have to show that the BJP 's victory in the 2014 Lok Sabha elections was a [...continues]	KGW: We have to study the ground reality and political situation in the city. We can't ignore the sweeping victory of the BJP in the 2014 elections in Delhi. The party won all the seven parliamentary seats. [...continues]	KGW: JP participated in the 2014 Lok Sabha elections in Delhi, where the party won just two seats. "We need to take into account the political climate and the political situation in the city. We cannot support the BJP's performance in the 2014 Lok Sabha elections [...continues]
... Daniels, who was born Stephanie Clifford, was paid 130,000 by Cohen after she signed a nondisclosure agreement barring her from talking about her alleg...	EXP: ed affair with Trump. Trump denies the allegation. Cohen has admitted making the payment to Daniels, which he said was done to protect Trump' s campaign from the allegations. It' s possible that the reimbursement payment revealed in the financial disclosure may have been [...continues]	EXP: ed affair with Trump. Cohen recently revealed that Trump personally reimbursed him for the payment to Daniels. As a result of the payment to Cohen, the government ethics office sent a reminder to Trump that he must disclose in his annual financial disclosure report [...continues]	EXP: The company announced that it would pay \$ 50,000 for the silence of this alleged affair. Cohen admitted paying the money, saying it was in order to protect Trump's campaign from the alleged affair. Trump denies the alleged affair. [...continues]	EXP: daniels spoke with Trump. trump denied the allegation. Cohen later admitted to the payments to Daniels, which he said were done to protect Trump's family from the allegations. It's possible that the cash payments mentioned in the full disclosure would [...continues]
... other shows, it's a GoPro on a windshield, " Foley said referring to Ride Along. " I think if this was a show that was on...	SIR: a traditional television platform, we would be able to do more with it. As it stands, it 'll be a while before we do any new episodes of Holy Folesy! " Foley also addressed why his daughter Noelle is n't pursuing a career in WWE despite [...continues]	SIR: a major network television, we would be able to do more with it. I think we would be able to have a bigger budget and be able to do some cool things with it. " Foley also talked about why his daughter Noelle is n't pursuing an [...continues]	SIR: Furthermore, Foley was asked why his daughter Nol was not continuing in the world of professional wrestling, despite her father's rich career. It's too early for her to become a wrestler, because the thing is that it's [...continues]	SIR: a great wrestling platform, we would be happy to do something with it, as it stands, it'll be a while before we have any more fans of holy Folesy! " Foley ##a explained why his daughter Noelle isn't pursuing a career in WWE despite. [...continues]

Appendix D. Ablation

To better understand the effect of noise injection location, we conduct a study and report results in Table 8 comparing three strategies on LLaMA-2 with KGW at $\delta = 0.01$ on (1) noisy embeddings during fine-tuning (δ -STEAL), (2) noisy pre-logits, and (3) noisy embeddings applied at inference.

Noise Location	PPL	AttackSR (%)
Noisy Embedding (training, δ -STEAL)	4.30	68.33%
Noisy Pre-logits	4.86	74.75%
Noisy Embedding (inference)	5.27	67.69%

Table 8: Ablation study comparing noise injection locations on LLaMA-2 with KGW at $\delta = 0.01$.