A Appendix A: Proof of Theorem 1

Let $\mathcal{A}$ be a randomized mechanism, R_θ be model parameter space, ϵ denote privacy budget, and σ denote the noise associated with ϵ.

In ϵ-STEAL, we add an LDP noise $\mathcal{N}(0, \sigma)$ to the token embeddings, it is equivalent to add ϵ-LDP to each data sample. By the post-processing property of LDP, the adversary model θ_{adv} is also ϵ-LDP. Therefore, IP checkers cannot tell whether θ_{adv} was trained on the watermark data or not and fail to verify the IP of θ_{adv}. Or the difference between θ_{adv} is trained with and without watermarked data is bounded, as follows:

$$\begin{aligned} &P[\mathcal{A}(x + \mathcal{N}(0,\sigma), y^{wm} + \mathcal{N}(0,\sigma)) \in R_\theta] \\ \leq &\exp^{\epsilon} P[\mathcal{A}(x' + \mathcal{N}(0,\sigma), y'^{wm} + \mathcal{N}(0,\sigma)) = R_\theta] \end{aligned} \tag{3}$$

Therefore, Theorem 1 holds.

B Appendix B: Privacy Budget Calculation

In our ϵ-STEAL, we use a common LDP approach, which is a Laplace mechanism that adds Laplace noises into original embeddings of the model. The Laplace mechanism is defined as follows:

$$\mathcal{A}_{\mathcal{E}}(x, \mathcal{E}(x), \epsilon) = \mathcal{E}(x) + (L_1, L_2, \cdots, L_d) \tag{4}$$

where $\mathcal{E}(x)$ is an embedding of a token x, d is the size of embedding, and L_i is i.i.d. random variables draw from a Laplace noise that is centered at 0 (i.e., mean is 0) and is scaled with $\sigma = \frac{\Delta(\mathcal{E})}{\epsilon}$.

Given a noise scale σ, to compute the privacy budget ϵ, we need to compute $\Delta(\mathcal{E})$, as follows:

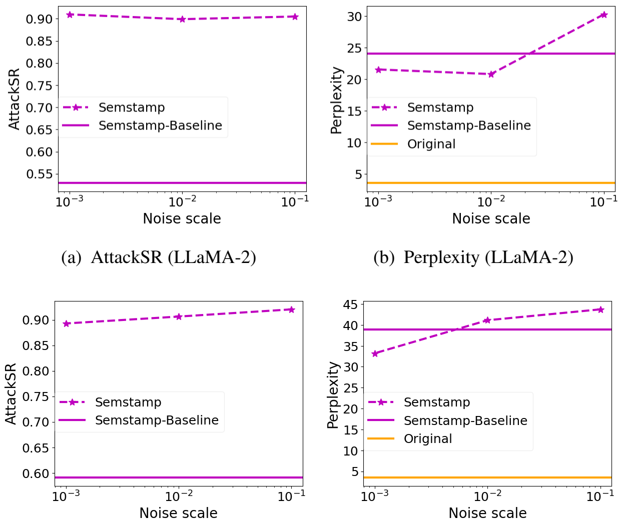
$$\Delta(\mathcal{E}) = \max_{\forall x, \tilde{x} \in N^d} \|\mathcal{E}(x) - \mathcal{E}(\tilde{x})\|_1 \tag{5}$$

With the LLaMA-2, we obtain $\Delta(\mathcal{E}) = 0.3$, while with the Mistral, we obtain $\Delta(\mathcal{E}) = 0.05$.

Then, with noise scales of $\sigma \in \{0.001, 0.01, 0.05, 0.1\}$, the privacy budgets are $\epsilon = \frac{\Delta(\mathcal{E})}{\delta} = \{300, 30, 6, 3\}$ with the LLaMA-2 and $\{50, 5, 1, 0.5\}$ for the Mistral.

C Appendix C: Supplemental Results

ϵ-STEAL against Semstamp.



(a) AttackSR (LLaMA-2) (b) Perplexity (LLaMA-2)

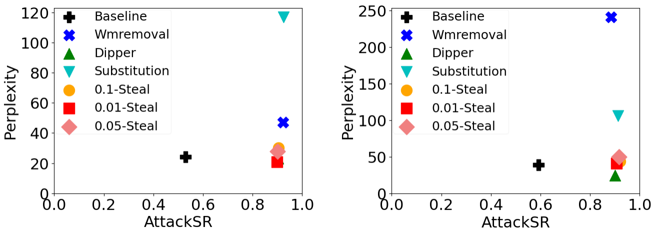
(c) AttackSR (Mistral) (d) Perplexity (Mistral)

Fig. 6: AttackSR and PPL of ϵ-STEAL on Semstamp and two LLMs.

For Semstamp, our observations of ϵ-STEAL attacks in Fig. 6 remain consistent with other WMs. As noise scales increase, the attack success rates also increase, while PPLs rise but remain comparable to the Baseline. However, a notable concern is the low Baseline WM detection rate, which results in unexpectedly high Attack SR even without attacks—reaching 52.97% for LLaMA-2 and 59.15% for Mistral.

ϵ-STEAL and Existing Model Stealing Attacks on Semstamp.

Fig. 7 demonstrates the effectiveness of our ϵ-STEAL attacks compared with other attacks on Semstamp. It is evident that ϵ-STEAL regardless of noise scales performs effectively on Semstamp as they gather at the bottom right corner, signifying high AttackSR but low PPLs.



(a) LLaMA-2 with Semstamp (b) Mistral with Semstamp

Fig. 7: Attack SR and Perplexity across different LLMs and Attacks including ϵ-STEAL on Semstamp.

ϵ-STEAL in MMLU Downstream Task.

In Table 1, we present only KGW and EXP watermarking techniques, although we have experimented with all the four watermarking techniques considered in this paper, including SIR and Semstamp. With the SIR and Semstamp watermarking techniques, our attempts to apply SIR watermark for MMLU task using its publicly available trained watermark model resulted in a poor accuracy, only 27.90% for the LLaMA-2 and 20.90% for the Mistral without any attacks. The root cause of such poor results is because due to time and computational power limitation, we do not retrain their watermark model, therefore it may not adapt well to our data and settings. For Semstamp, since it generates sentences instead of single token at once, it is considered not suitable for the MMLU, which involves question-answering tasks.

Semantic Preservation of ϵ-STEAL Outputs.

In the following section, we present additional examples from the Mistral, as shown in Table 3a. These examples demonstrate the outputs of our ϵ-STEAL attack, highlighting the differences while preserving the overall semantic meaning. For example, in the first row, the attack replaces“office has talked to Attorney General Jeff Session” with “team has talked with Jeff Sessions” or remove adverb “now”, 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 school” shifts to“who applied to the city’s top schools were admitted” subtly altering the focus while retaining the core message that the mentioned students still got in top schools. Lastly, in the third instance, ϵ-STEAL modifies“shape of a new toy” to “shape, configuration, and appearance of a new product,” and changes the quantifier from plural to singular, “patent” and “variety”, preserving the discussion on patent types. These examples confirm that our attack slightly modifies wording without compromising the overall semantic content.

Table 3b provides a comparison of our ϵ-STEAL with the different attack methods used in this study. While the approach of our ϵ-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 ϵ-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 3b 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 (more green portion) with the original watermarked text.

Table 3a: Selected prompts and watermarked outputs with and without ϵ-STEAL attack. The examples are from Mistral model under the KGW (the second row) and SIR (the first and third row) watermarks. Green texts indicate similarities, while red texts highlight differences.

Prompt	**Watermarked (W)**	**Watermarked under ϵ-Steal (SA)** ($\epsilon = 0.01$)	**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 ...	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]	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-...	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]	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...	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]	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 3b: Output comparison among ϵ-STEAL and other attacks for the same input. The examples are from LLaMA-2 model under the KGW (the first row), EXP (the second row), and SIR (the third row) watermarks. Green texts indicate similarities, while red texts highlight differences.

Prompt	**Watermarked (W)**	**Watermarked under ϵ-Steal (SA)** ($\epsilon = 0.01$)	**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 . . .	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]	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]	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]	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...	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]	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]	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]	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...	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]	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]	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]	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]