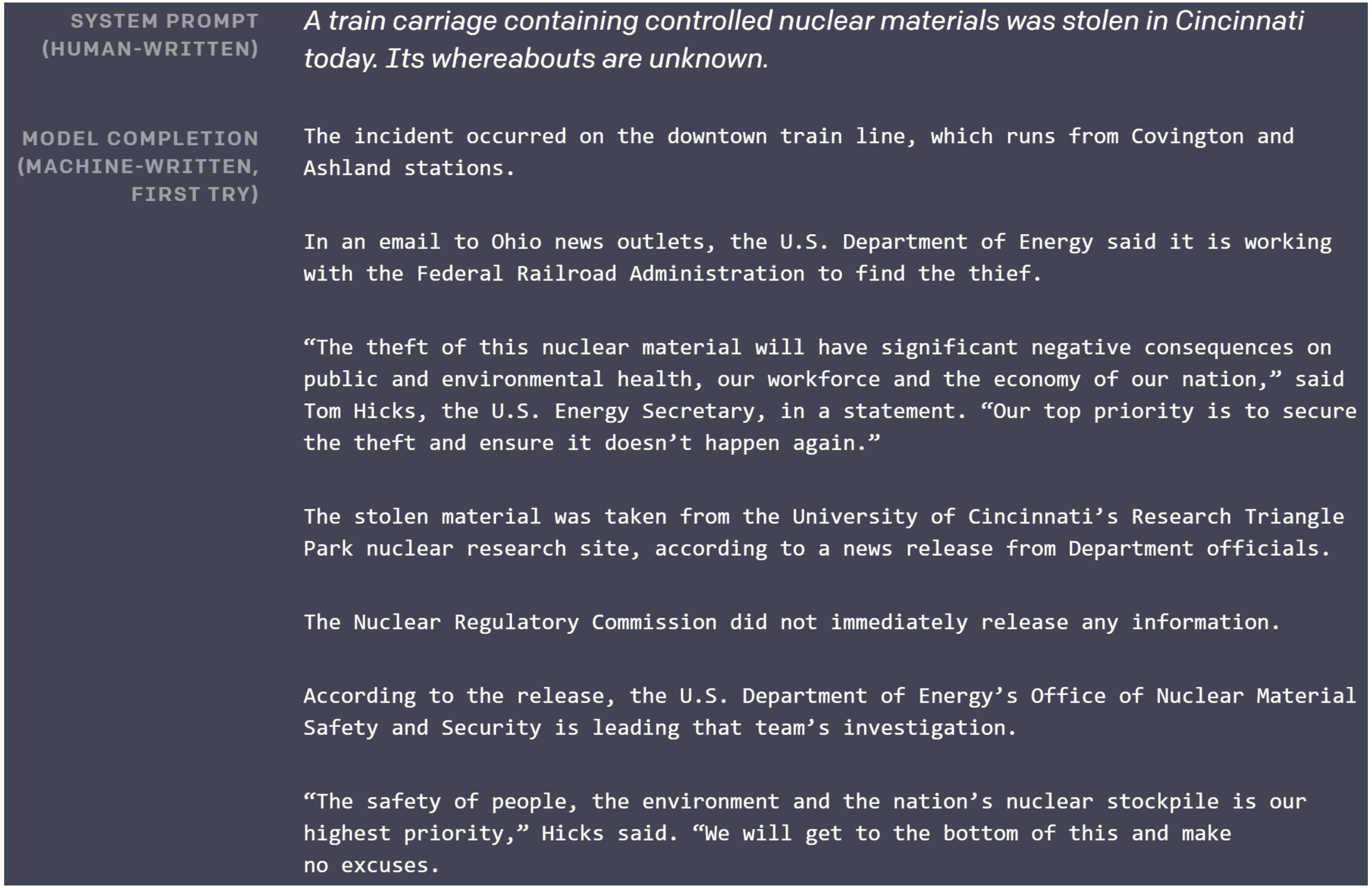
Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding

Sahar Abdelnabi and Mario Fritz
CISPA Helmholtz Center for Information Security

Motivation

Recent language models create realistically-looking text



SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

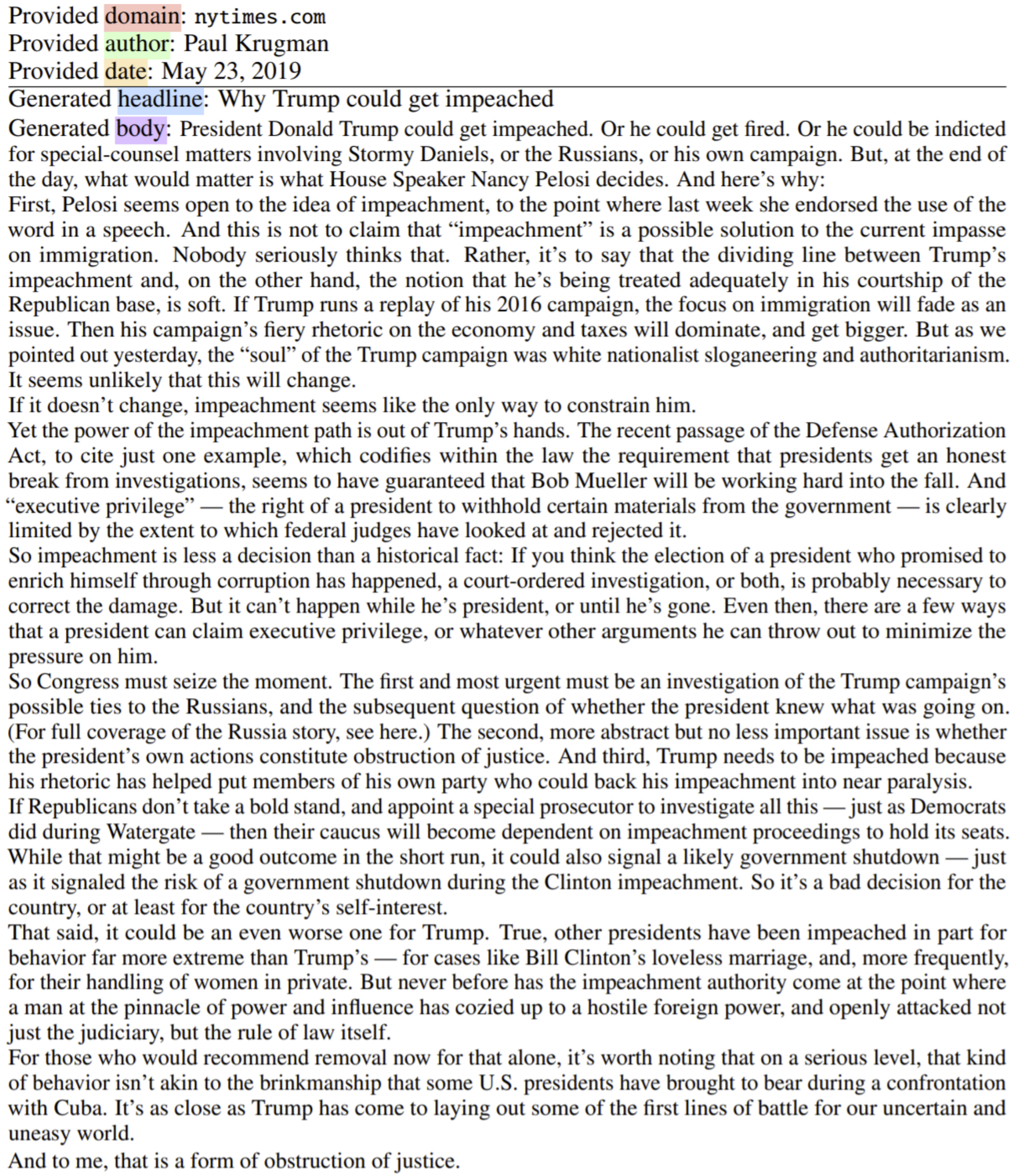
According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

Example of GPT-2 from: https://openai.com/blog/better-language-models

Motivation

Recent language models create realistically-looking text



Provided domain: nytimes.com
Provided author: Paul Krugman
Provided date: May 23, 2019

Generated headline: Why Trump could get impeached

Generated body: President Donald Trump could get impeached. Or he could get fired. Or he could be indicted for special-counsel matters involving Stormy Daniels, or the Russians, or his own campaign. But, at the end of the day, what would matter is what House Speaker Nancy Pelosi decides. And here's why:

First, Pelosi seems open to the idea of impeachment, to the point where last week she endorsed the use of the word in a speech. And this is not to claim that "impeachment" is a possible solution to the current impasse on immigration. Nobody seriously thinks that. Rather, it's to say that the dividing line between Trump's impeachment and, on the other hand, the notion that he's being treated adequately in his courtship of the Republican base, is soft. If Trump runs a replay of his 2016 campaign, the focus on immigration will fade as an issue. Then his campaign's fiery rhetoric on the economy and taxes will dominate, and get bigger. But as we pointed out yesterday, the "soul" of the Trump campaign was white nationalist sloganeering and authoritarianism. It seems unlikely that this will change.

If it doesn't change, impeachment seems like the only way to constrain him.

Yet the power of the impeachment path is out of Trump's hands. The recent passage of the Defense Authorization Act, to cite just one example, which codifies within the law the requirement that presidents get an honest break from investigations, seems to have guaranteed that Bob Mueller will be working hard into the fall. And "executive privilege" — the right of a president to withhold certain materials from the government — is clearly limited by the extent to which federal judges have looked at and rejected it.

So impeachment is less a decision than a historical fact: If you think the election of a president who promised to enrich himself through corruption has happened, a court-ordered investigation, or both, is probably necessary to correct the damage. But it can't happen while he's president, or until he's gone. Even then, there are a few ways that a president can claim executive privilege, or whatever other arguments he can throw out to minimize the pressure on him.

So Congress must seize the moment. The first and most urgent must be an investigation of the Trump campaign's possible ties to the Russians, and the subsequent question of whether the president knew what was going on. (For full coverage of the Russia story, see here.) The second, more abstract but no less important issue is whether the president's own actions constitute obstruction of justice. And third, Trump needs to be impeached because his rhetoric has helped put members of his own party who could back his impeachment into near paralysis.

If Republicans don't take a bold stand, and appoint a special prosecutor to investigate all this — just as Democrats did during Watergate — then their caucus will become dependent on impeachment proceedings to hold its seats. While that might be a good outcome in the short run, it could also signal a likely government shutdown — just as it signaled the risk of a government shutdown during the Clinton impeachment. So it's a bad decision for the country, or at least for the country's self-interest.

That said, it could be an even worse one for Trump. True, other presidents have been impeached in part for behavior far more extreme than Trump's — for cases like Bill Clinton's loveless marriage, and, more frequently, for their handling of women in private. But never before has the impeachment authority come at the point where a man at the pinnacle of power and influence has cozied up to a hostile foreign power, and openly attacked not just the judiciary, but the rule of law itself.

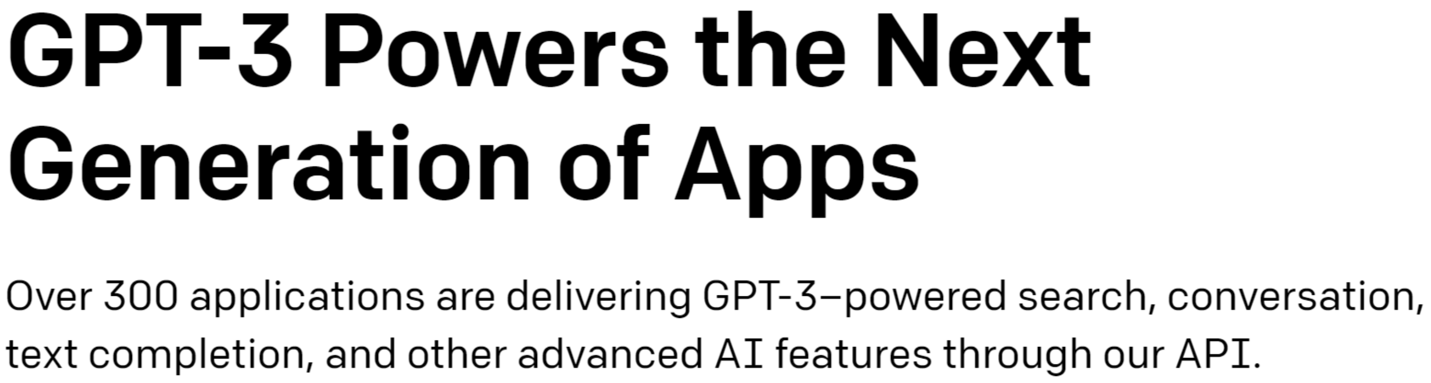
For those who would recommend removal now for that alone, it's worth noting that on a serious level, that kind of behavior isn't akin to the brinkmanship that some U.S. presidents have brought to bear during a confrontation with Cuba. It's as close as Trump has come to laying out some of the first lines of battle for our uncertain and uneasy world.

And to me, that is a form of obstruction of justice.

Example of Grover from: Zellers et al. "Defending Against Neural Fake News." *NeurIPS* (2020).

Motivation

Recent language models create realistically-looking text



GPT-3 Powers the Next Generation of Apps

Over 300 applications are delivering GPT-3–powered search, conversation, text completion, and other advanced AI features through our API.

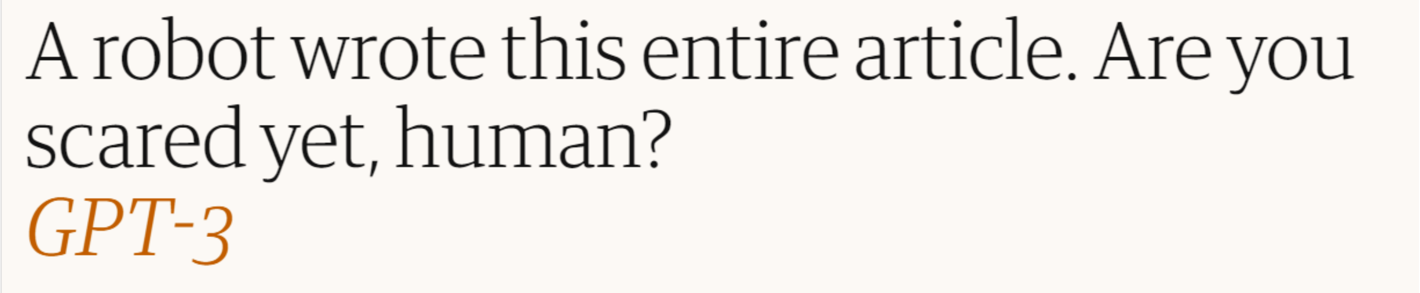
https://openai.com/blog/gpt-3-apps/

Motivation

Recent language models create realistically-looking text

Concerns about the implications of AI technologies

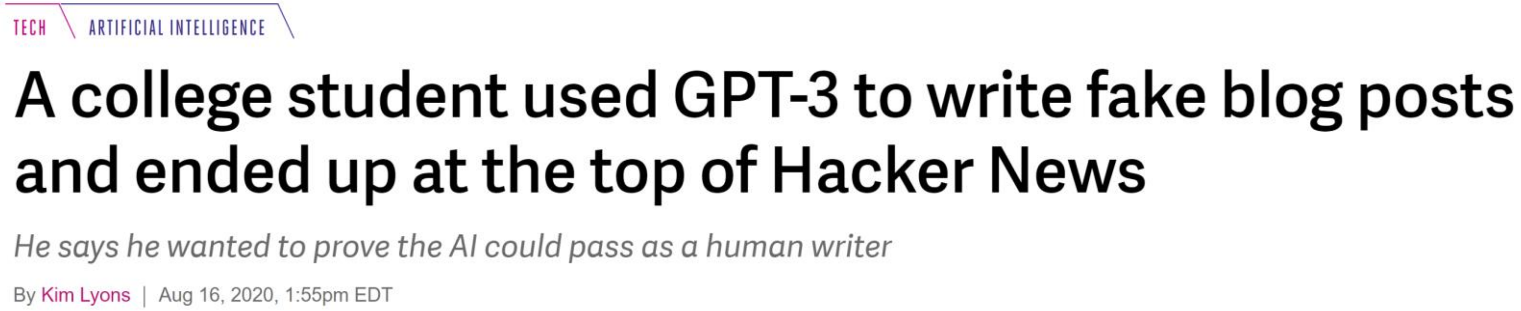
Counter release strategies (staged release, black-box APIs)



A robot wrote this entire article. Are you scared yet, human?

GPT-3

The guardian (https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3).



TECH ARTIFICIAL INTELLIGENCE

A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

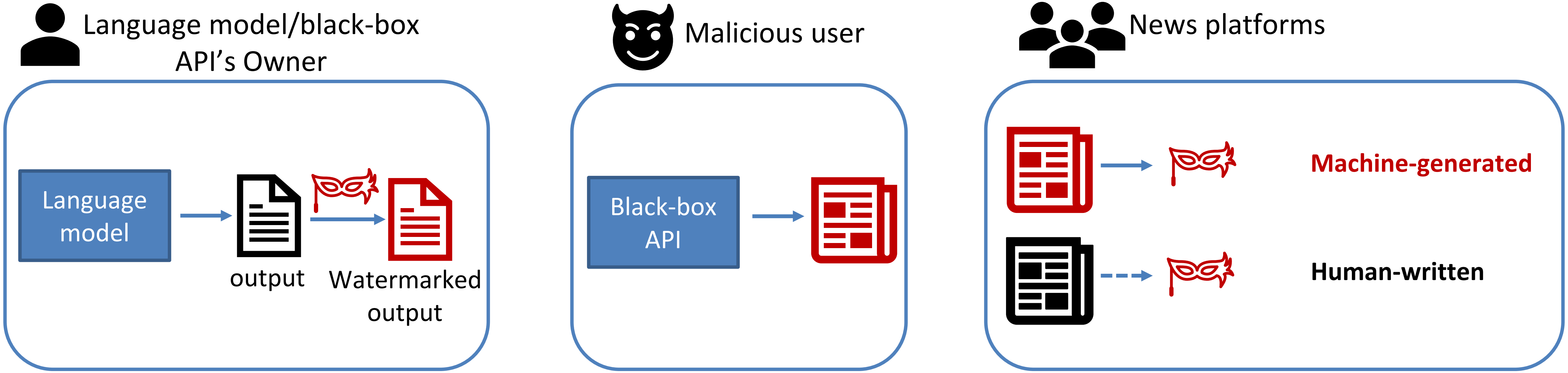
He says he wanted to prove the AI could pass as a human writer

By Kim Lyons | Aug 16, 2020, 1:55pm EDT

https://www.theverge.com/2020/8/16/21371049/gpt3-hacker-news-ai-blog

Towards tracing text provenance

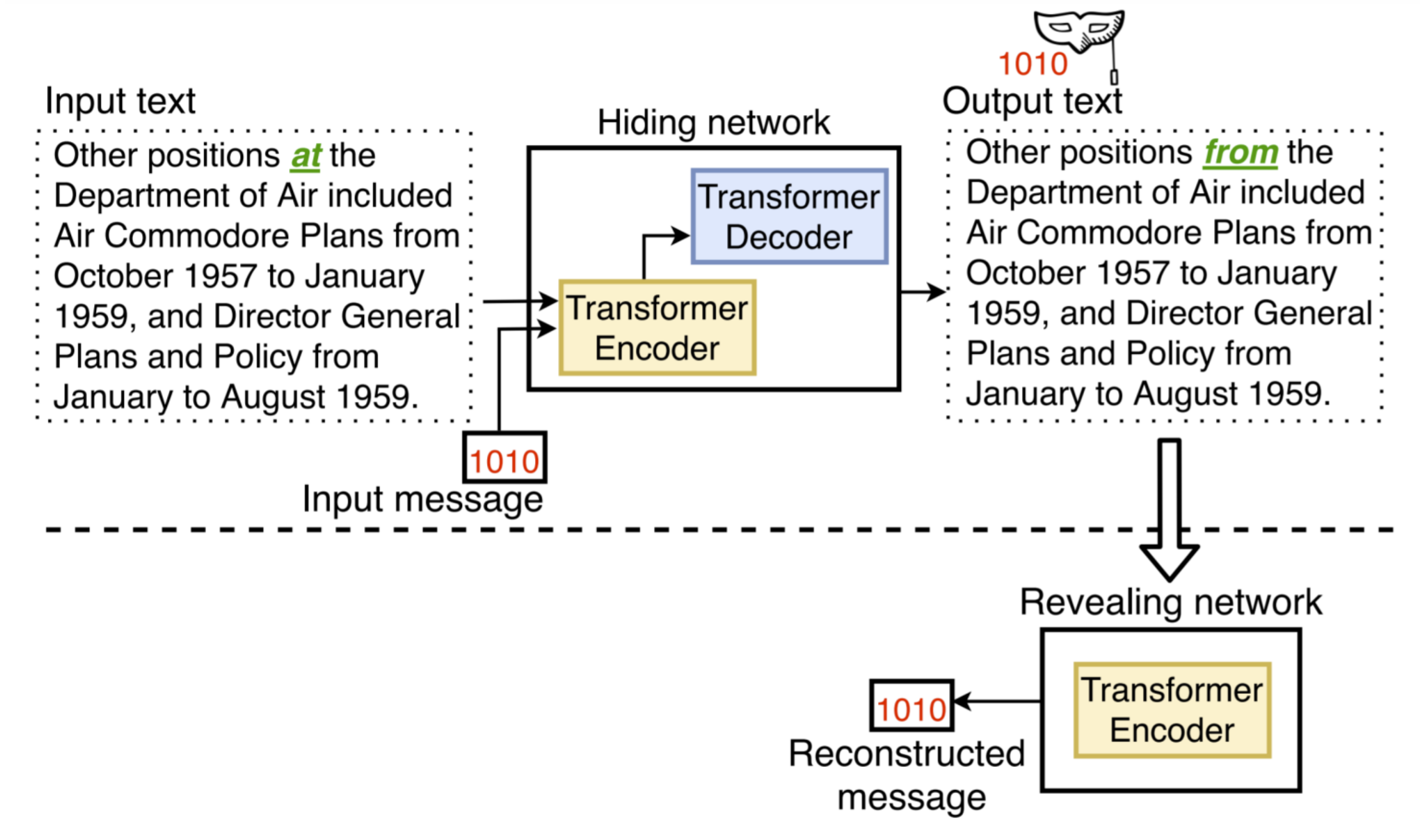
Moving from ***passive*** to ***active*** defenses against deepfakes



Towards responsible release and regulation of AI models

Our approach: ***Adversarial Watermarking Transformer (AWT)***

AWT is the first end-to-end framework for unobstructively watermarking text


1010
Output text
Other positions *from* the Department of Air included Air Commodore Plans from October 1957 to January 1959, and Director General Plans and Policy from January to August 1959.
Input text
Other positions *at* the Department of Air included Air Commodore Plans from October 1957 to January 1959, and Director General Plans and Policy from January to August 1959.
Hiding network
Transformer Decoder
Transformer Encoder
1010
Input message
Revealing network
Transformer Encoder
1010
Reconstructed message

Evaluation axes

Effectiveness ★★★

Output **quality** in relation with message **bit accuracy**

Secrecy

Watermarks should be ***indistinguishable***

Robustness

Resilience against automated removal attempts

These evaluation axes might be *competing* and reaching a *trade-off* is required

Related work

Rule-based methods

- Designed rules
- Large changes, lower utility
- Fixed changes
- Could compromise secrecy and robustness

AWT

- Automatic and learned
- Subtle, minimal changes
- More flexibility
- Better secrecy and robustness

Topkara et al. “Natural language watermarking”, 2005.

Topkara et al. “Natural language watermarking: Challenges in building a practical system”, 2006.

Topkara et al., “The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions”, 2006.

Meral et al. “Natural language watermarking via morphosyntactic alterations”, 2009.

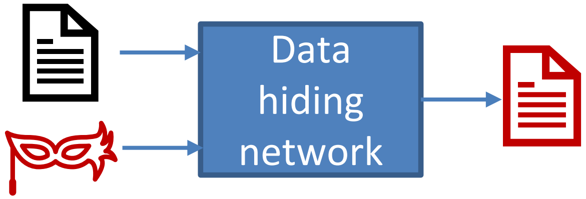
Wilson et al. “Linguistic steganography on twitter: hierarchical language modeling with manual interaction”, 2014.

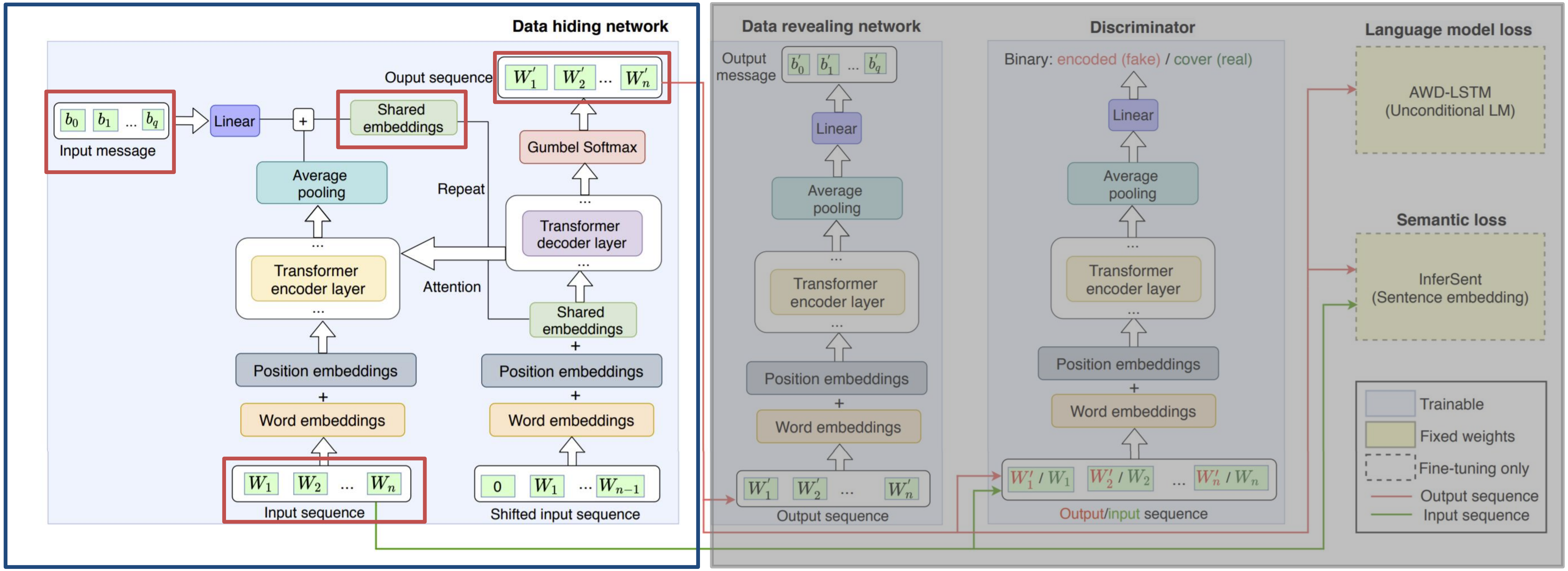
Shirali-Shahreza et al. “A new synonym text steganography”, 2008.

Zhu et al. “Hidden: Hiding data with deep networks”, 2018.

Adversarial Watermarking Transformer (AWT)

- **Data-hiding network (Message encoder)**
  - A sequence-to-sequence transformer-based encoder and decoder



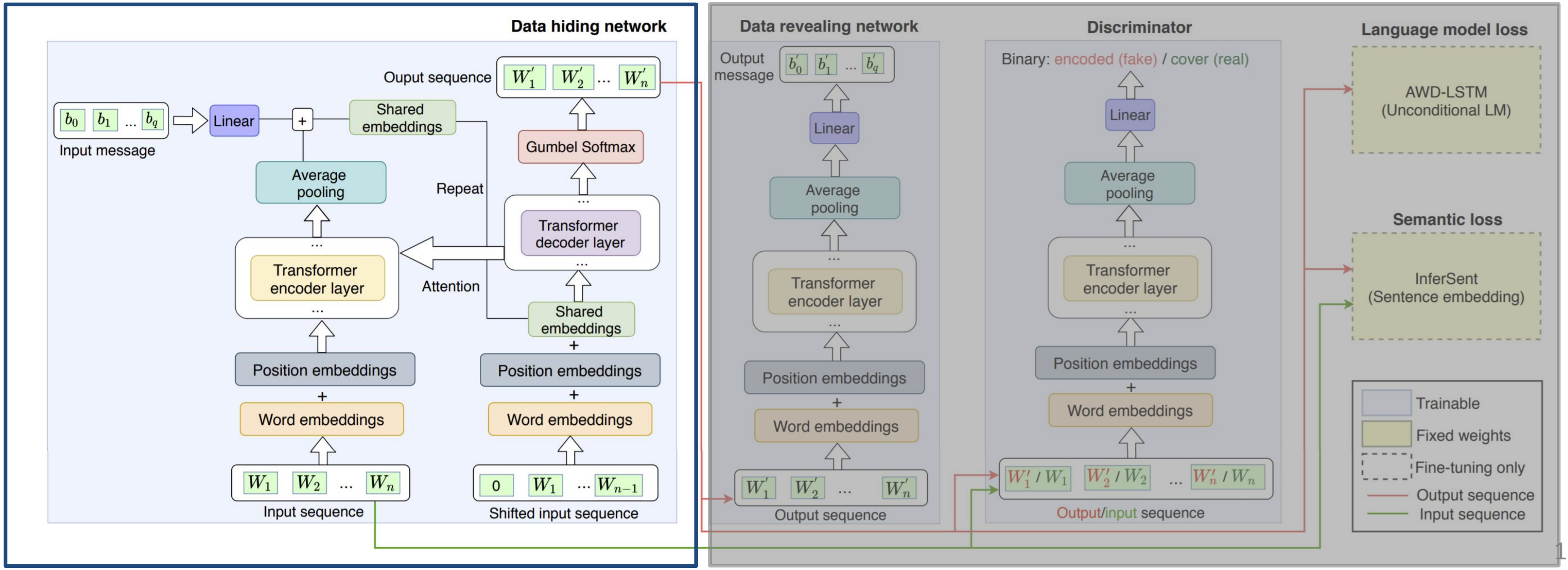


Adversarial Watermarking Transformer (AWT)

- **Data-hiding network (Message encoder)**
  - A sequence-to-sequence transformer-based encoder and decoder

  - The network is trained to reconstruct the input sequence (i.e., as an autoencoder)

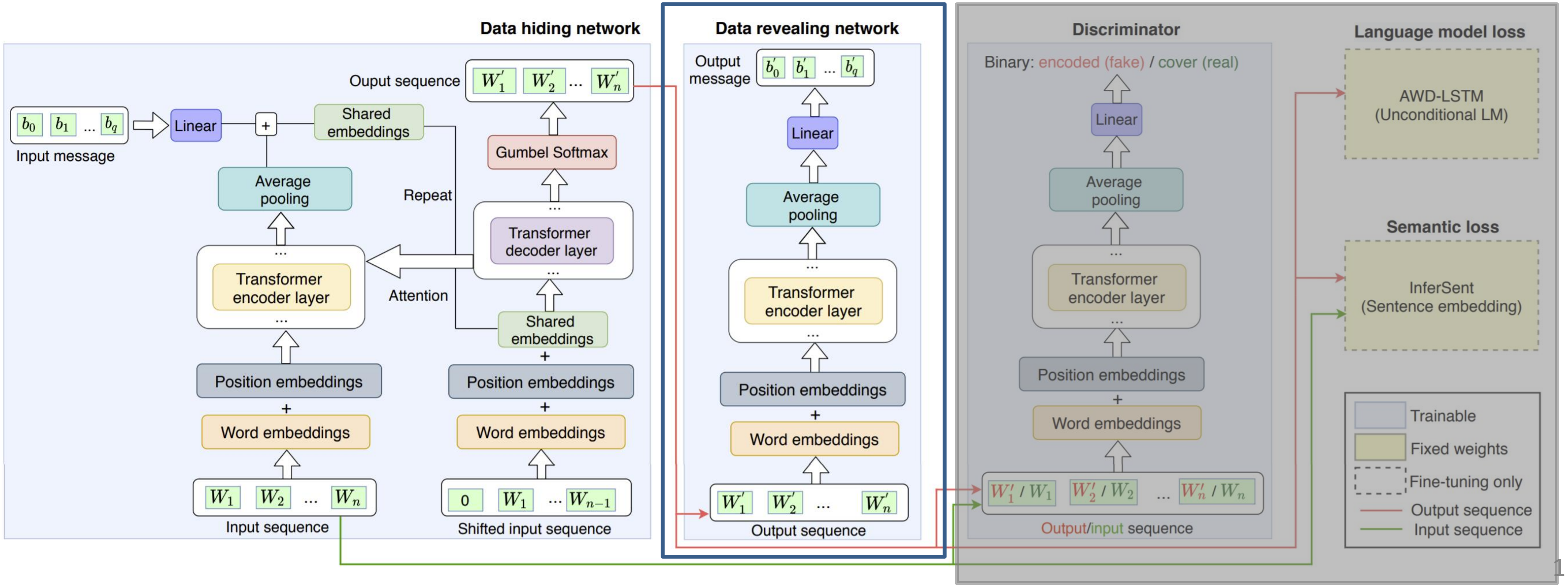
$$L_{rec} = \mathrm{E}_{p_{data}(S)}[-\log p_D(S)]$$



Adversarial Watermarking Transformer (AWT)

- **Data-revealing network (Message decoder)**
  - Takes the watermarked text and reconstructs the message

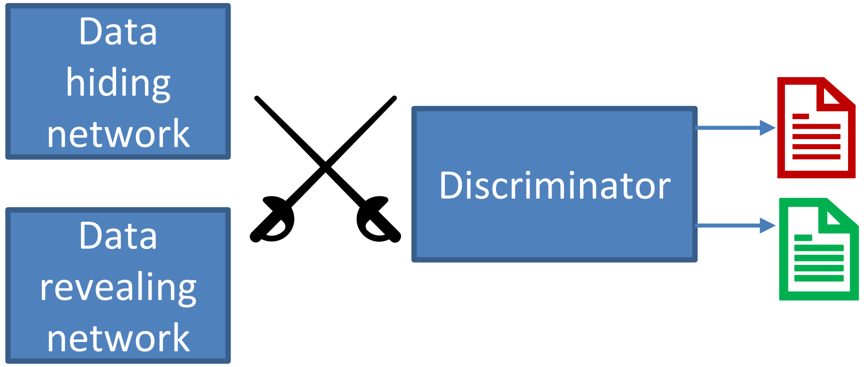
$$L_m = \sum_{i=1}^{q} b_i \log\big(p_M(b_i)\big) + (1 - b_i)\log(1 - P_M(b_i))$$

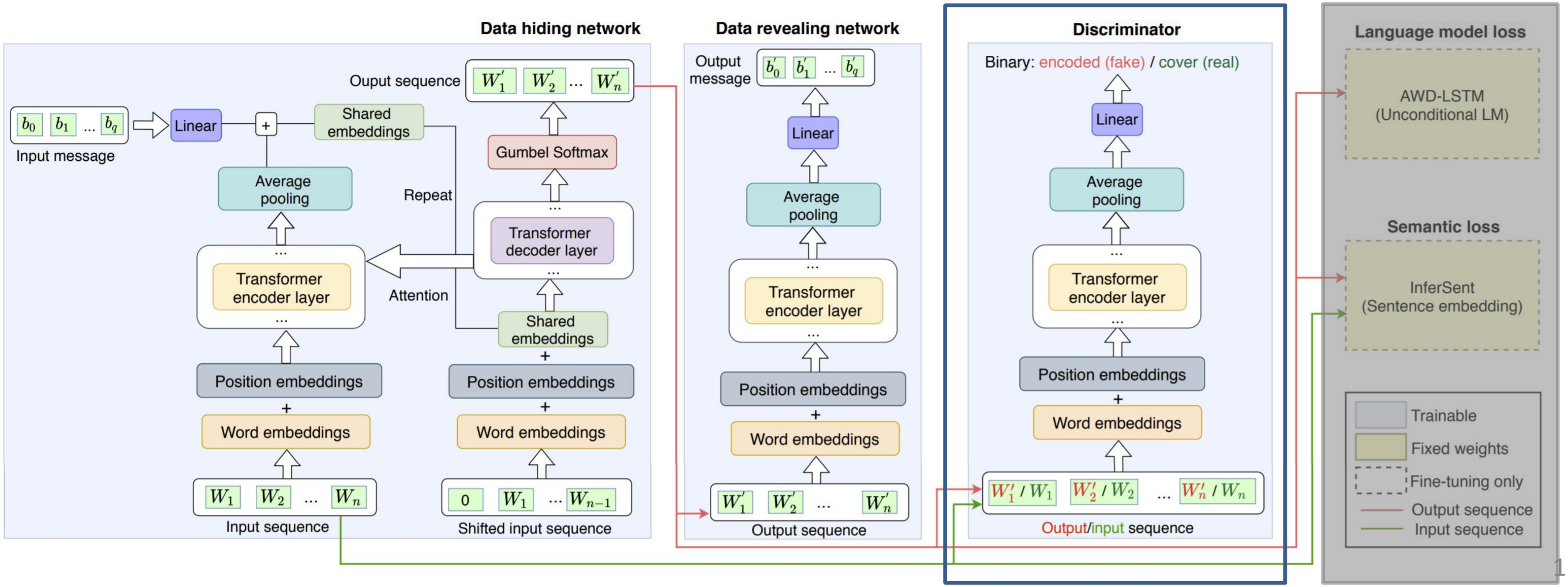


Adversarial Watermarking Transformer (AWT)

- **Discriminator (A)**
  - We utilize adversarial training against the previous components

$$L_A = -\log A(S^{`})$$

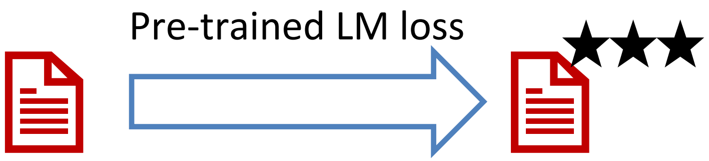




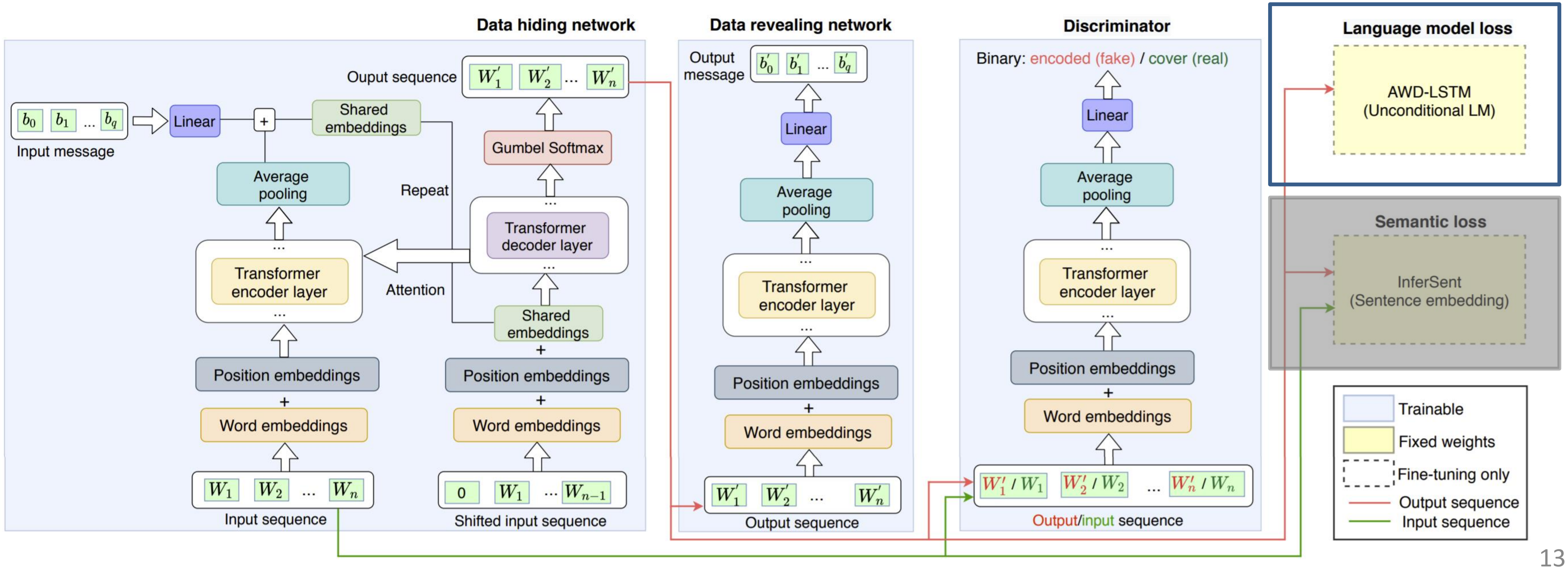
Adversarial Watermarking Transformer (AWT)

- **Fine-tuning**
  - **Correctness:** Maximize the likelihood of the output sentence under a pre-trained language model

$$L_{LM} = -\sum_i \log p_{LM}\left(\grave{W}_i | \grave{W}_{<i}\right)$$

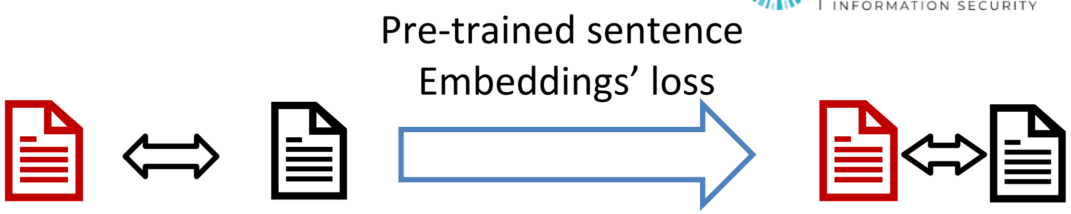


Merity et al., "Regularizing and optimizing lstm language models", 2018.



Adversarial Watermarking Transformer (AWT)

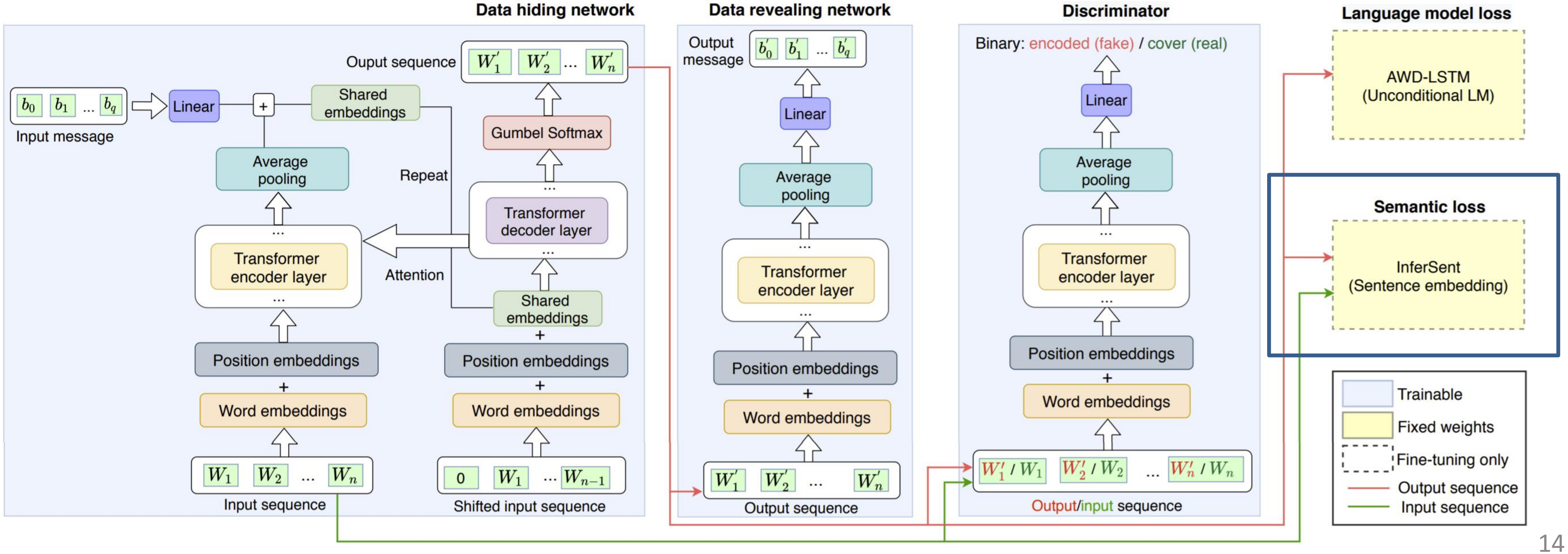
- **Fine-tuning**
  - **Semantic preserving:** Minimize the distance between the input and output sentences' embeddings



Pre-trained sentence Embeddings' loss

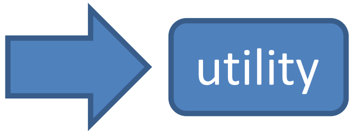
$$L_{sem} = \left\| F(S) - F(S^{`}) \right\|$$

Conneau et al., "Supervised learning of universal sentence representations from natural language inference data", 2017.



Experimental setup

- **Dataset**
  - The word-level WikiText-2 (WT2), curated from **Wikipedia** articles
- **Metrics**
  - **Meteor score** (higher is better)
  - **SBERT distance** (lower is better)
  - **Message bit accuracy** (random chance: 50%)



utility

Merity et al., “Pointer sentinel mixture models”, 2017.
Reimers et al. "Sentence-bert: Sentence embeddings using siamese bert-networks”, 2019.

Watermark embedding and verification

- Message length of **4 bits** per **sentence**
- Document-level by successive encoding
- The message are decoded from each segment
- Verification is done by **hypothesis testing** based on the number of **matching bit** (k) out of all bits (n)

$$\Pr(X > k|H_0) = \sum_{i=k}^{n} \binom{n}{i} 0.5^n$$

Encoded	1101	0010	1010	…	1111
Decoded	1100	0010	0110	…	1111
Matching	1100	0010	0110	…	1111

Effectiveness★★★ - Text utility

Effectiveness ★★★

Secrecy

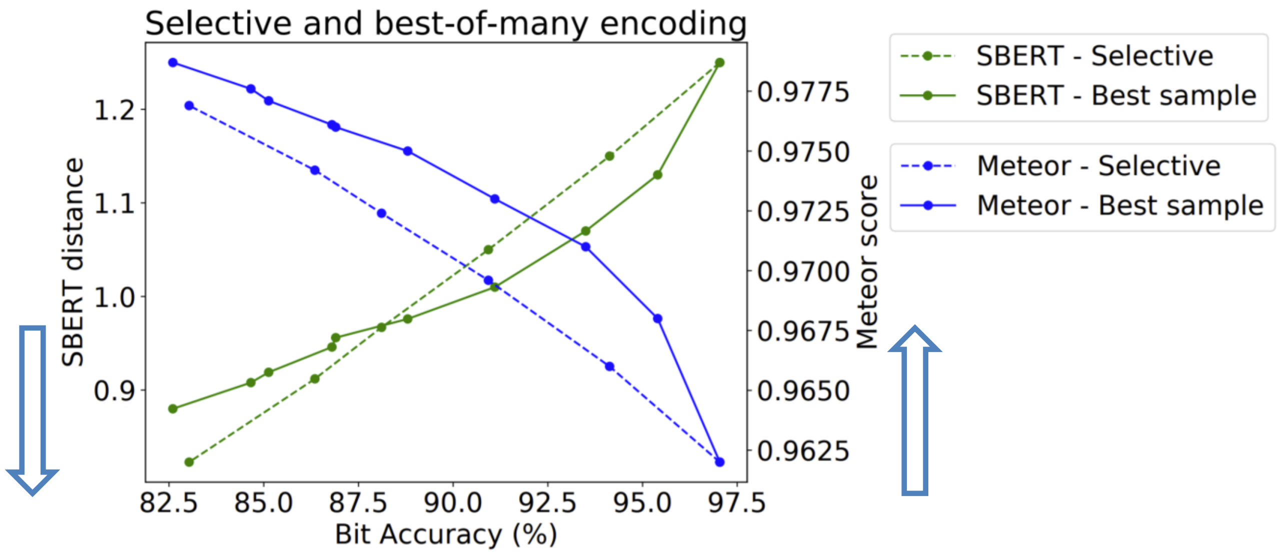
Robustness

- **Ablation**

Adversarial training and fine-tuning increase the utility

- **Improving text utility:**
  - **Sampling (best-of-many)**
    - Sample multiple output sentences
    - Select the best sample
  - **Selective encoding**
    - Leave some sentences unencoded

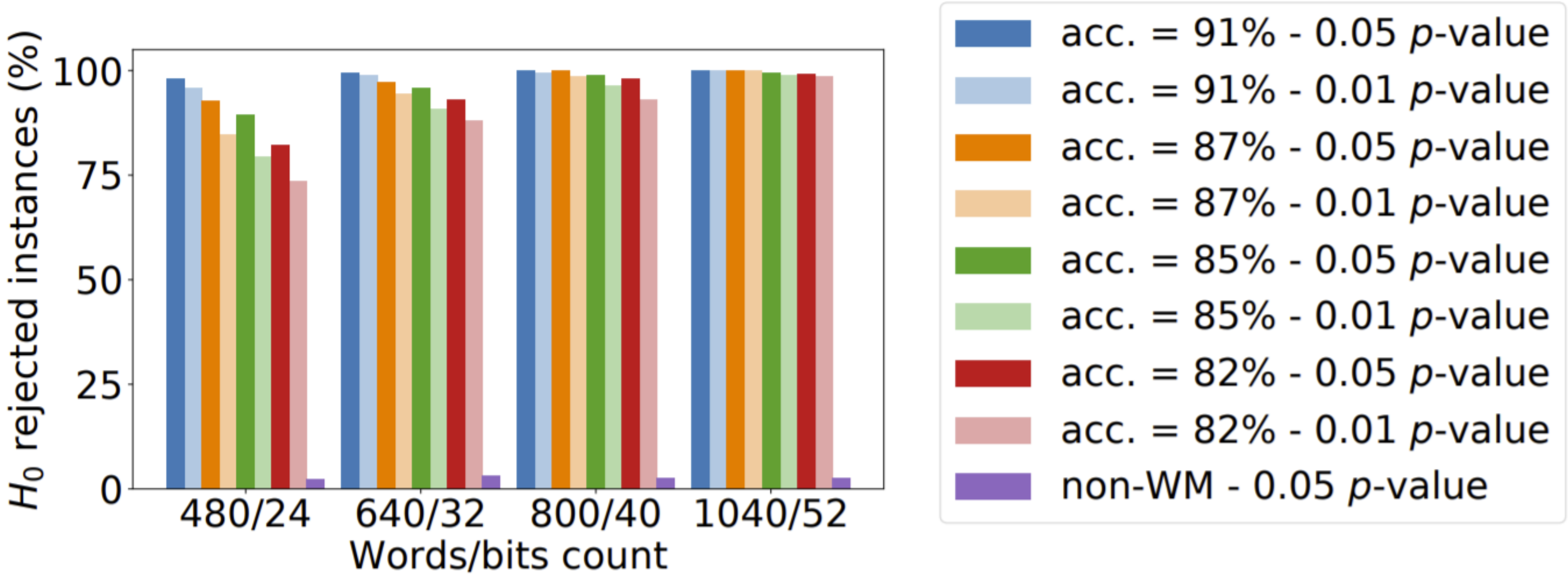
Variant	Bit accuracy ↑	Meteor ↑	SBERT↓
AWT	**97.04%±0.16**	**0.962±0.0003**	**1.26±0.008**
–fine-tuning	95.13%±0.21	0.943±0.0005	1.73±0.015
–discriminator	96.15%±0.22	0.938±0.0006	2.29±0.016



Effectiveness★★★ - Message decoding

- **Concatenate messages**
  - Verification by null hypothesis
  - Study detection rate vs. text length at different accuracy levels

Increasing text length increases detection rate



Effectiveness★★★ - Qualitative examples

- **“No discriminator” output**
  - Poor secrecy and naturalness

Input	Watermarked (no discriminator)
one of **the** most fascinating characters in **the** series	one of **Milton** most fascinating characters in **Milton** series
He was appointed **the** commanding officer.	He was appointed **Bunbury** commanding officer.

Effectiveness★★★ - Qualitative examples

- “No discriminator” output
  - Poor secrecy and naturalness
- “No fine-tuning” output
  - Uses likely tokens
  - Many mistakes

Input	Watermarked (no fine-tuning)
the Business Corporation, **which** was formed by a group of leaders from the area.	the Business Corporation, **<eos>** was formed by a group of leaders from the area.

Effectiveness★★★ - Qualitative examples

- **“No discriminator” output**
  - Poor secrecy and naturalness
- **“No fine-tuning” output**
  - Uses likely tokens
  - Many mistakes
- **Full *AWT***

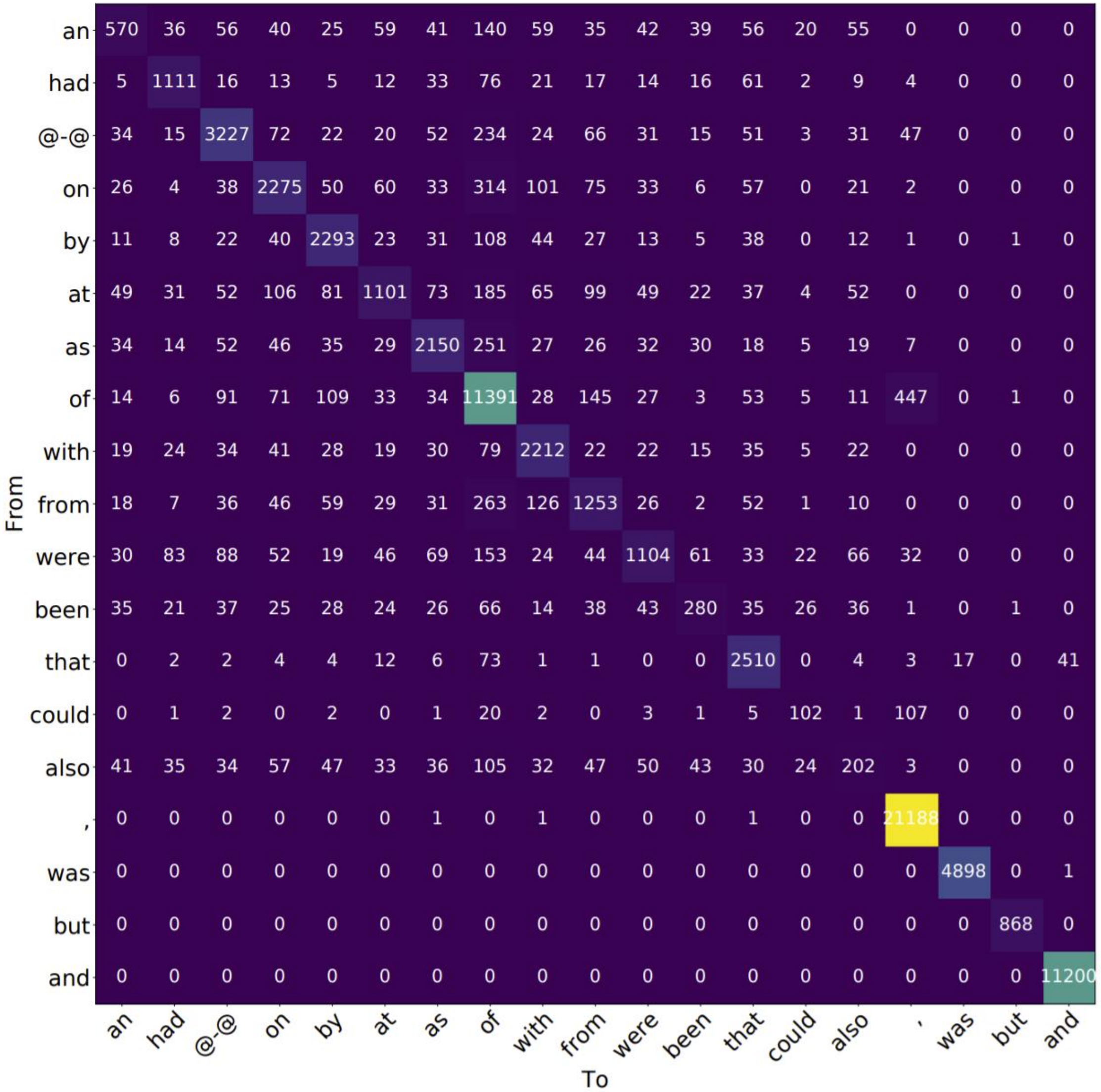
Better secrecy, correctness, and semantic coherency

Input	Watermarked (*AWT*)
In 1951 , a small airstrip was built **at** the ruins	In 1951 , a small airstrip was built **on** the ruins
It is the opening track **from** their 1987 album	It is the opening track **of** their 1987 album
, **but** the complex is broken up by the heat of cooking	, **and** the complex is broken up by the heat of cooking
He **also** performed as an actor and a singer	He **had** performed as an actor and a singer

Input	Watermarked (*AWT*)
government officials **had** been suspected	government officials **at** been suspected
who has **been** in office since 2009	who has **were** in office since 2009
The M @-@ 82 designation was truncated **at** this time	The M @-@ 82 designation was truncated **were** this time

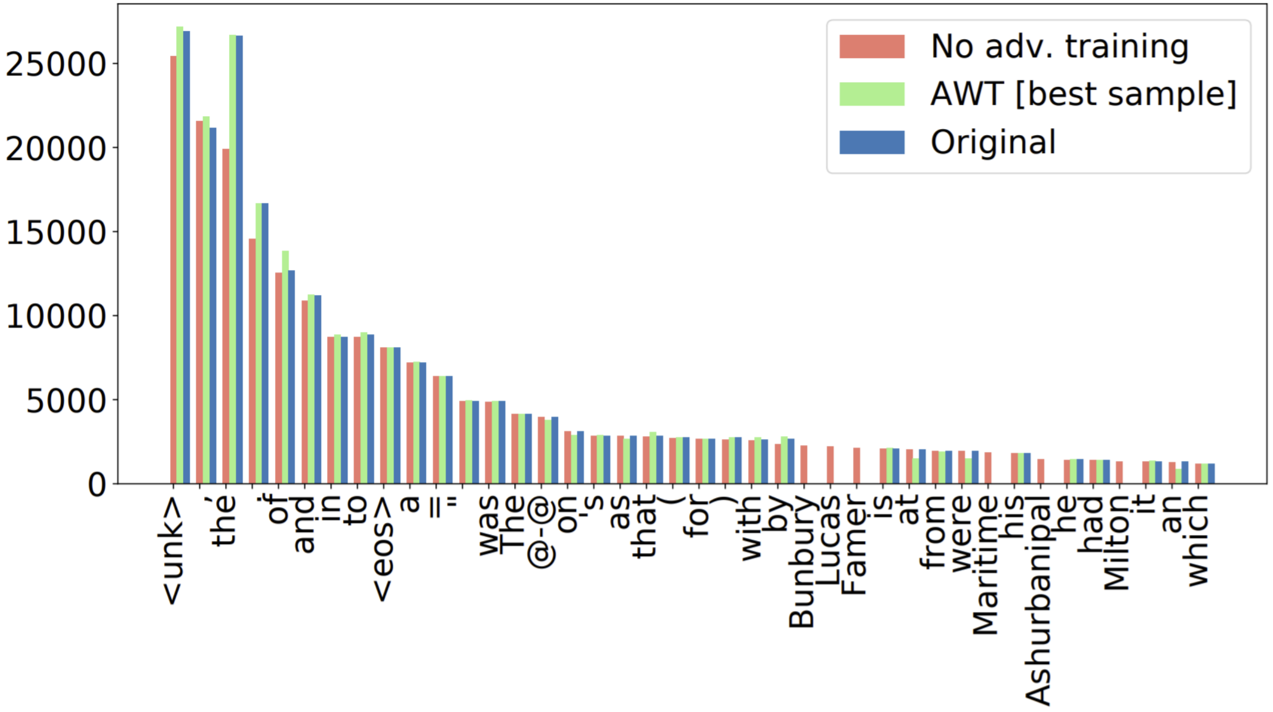
Secrecy - Analyzing word changes

- No fixed substitutions are observed



	an	had	@-@	on	by	at	as	of	with	from	were	been	that	could	also	,	was	but	and
an	570	36	56	40	25	59	41	140	59	35	42	39	56	20	55	0	0	0	0
had	5	1111	16	13	5	12	33	76	21	17	14	16	61	2	9	4	0	0	0
@-@	34	15	3227	72	22	20	52	234	24	66	31	15	51	3	31	47	0	0	0
on	26	4	38	2275	50	60	33	314	101	75	33	6	57	0	21	2	0	0	0
by	11	8	22	40	2293	23	31	108	44	27	13	5	38	0	12	1	0	1	0
at	49	31	52	106	81	1101	73	185	65	99	49	22	37	4	52	0	0	0	0
as	34	14	52	46	35	29	2150	251	27	26	32	30	18	5	19	7	0	0	0
of	14	6	91	71	109	33	34	11391	28	145	27	3	53	5	11	447	0	1	0
with	19	24	34	41	28	19	30	79	2212	22	22	15	35	5	22	0	0	0	0
from	18	7	36	46	59	29	31	263	126	1253	26	2	52	1	10	0	0	0	0
were	30	83	88	52	19	46	69	153	24	44	1104	61	33	22	66	32	0	0	0
been	35	21	37	25	28	24	26	66	14	38	43	280	35	26	36	1	0	1	0
that	0	2	2	4	4	12	6	73	1	1	0	0	2510	0	4	3	17	0	41
could	0	1	2	0	2	0	1	20	2	0	3	1	5	102	1	107	0	0	0
also	41	35	34	57	47	33	36	105	32	47	50	43	30	24	202	3	0	0	0
,	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	21188	0	0	0
was	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4898	0	1
but	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	868	0
and	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11200

- *AWT* keeps the top-words' count similar



Secrecy - Classifiers' performance

Secrecy

Effectiveness ★★★

Robustness

- Independent binary classifiers (watermarked vs non-watermarked) on non-paired sentences
- Lower F1 score indicates an unsuccessful attack

Model's variant	Sampling	Classifier	F1 score
– discriminator	1 sample	Transformer	0.89
		LSTM	0.8
		BoW+SVM	0.98
– fine-tuning	1 sample	Transformer	0.65
		LSTM	0.56
		BoW+SVM	0.63
	Best-of-20	Transformer	0.59
AWT	**Best-of-20**	**Transformer**	**0.53**

Improving the quality → higher secrecy

Robustness

Removal attacks aim at ***removing*** the watermark with ***minimal changes*** to the text

- **Trade-off** between:
  - The **decrease** of bit **accuracy**
  - The **increase of SBERT** w.r.t the original sentence

- Categorize the attacks w.r.t the adversary's knowledge:

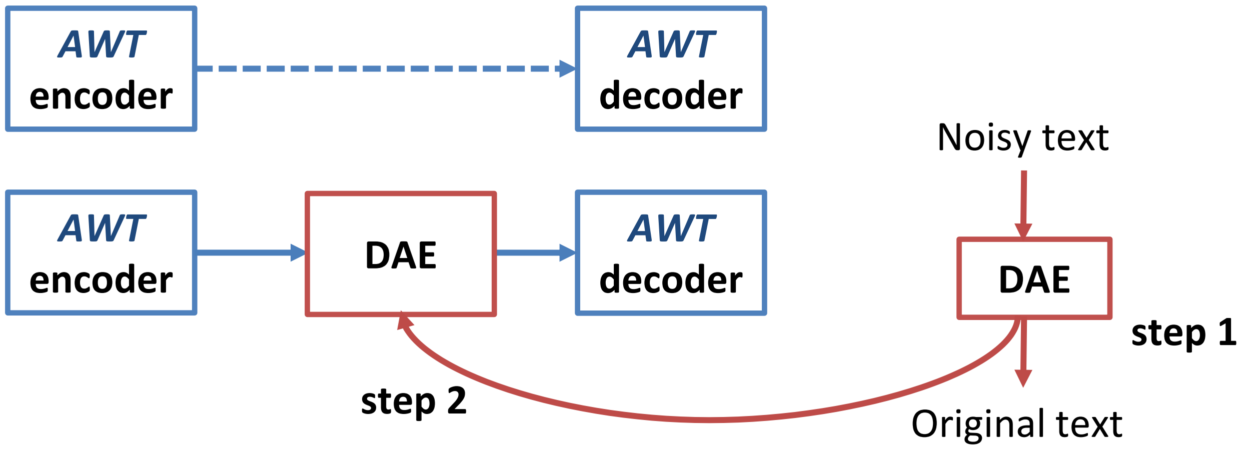
  1) **Basic attacks:** partial knowledge

  2) **Adaptive attacks:** full knowledge about ***AWT*** – but **without** white-box access

Robustness - Basic attacks

- **De-noising**
  - Train a de-noising autoencoder
  - Apply the de-noising autoencoder to watermarked text

Improving the quality → less successful denoising



AWT encoder

AWT decoder

AWT encoder

DAE

AWT decoder

Noisy text

DAE

step 1

Original text

step 2

Model	Bit accuracy drop	SBERT change
AWT	**1.93%±0.19**	30.77%±1.03↑
– fine-tuning	5.21%±0.12	14.20%±1.11↑
– discriminator	47.92%±0.44	15.93%±0.94↓

Robustness - Basic attacks

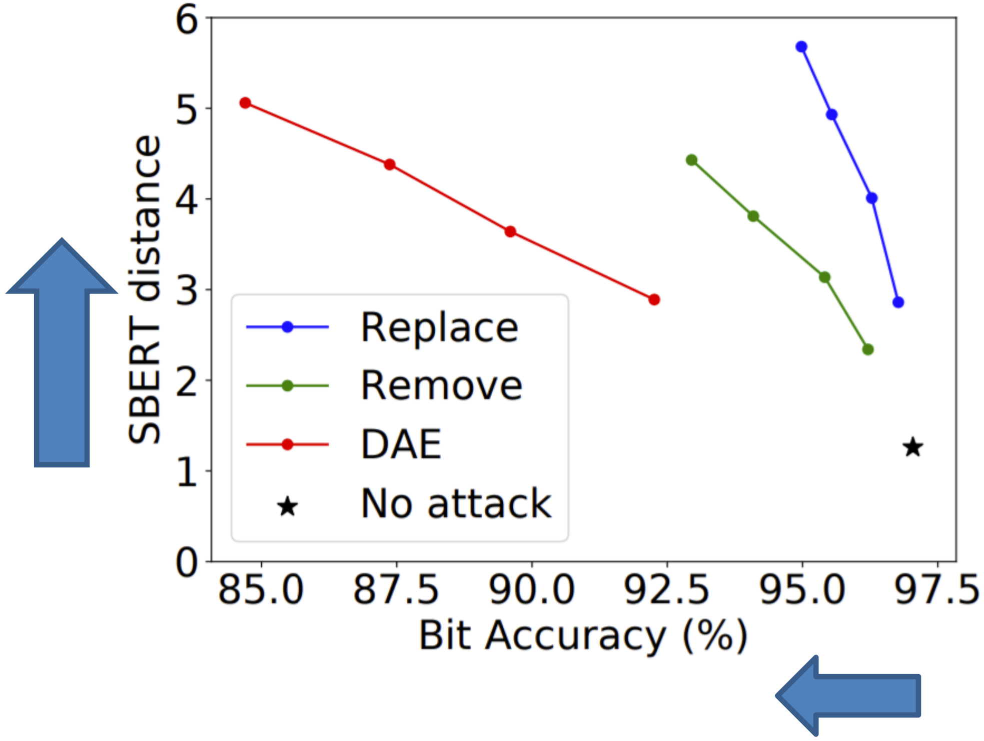
- **De-noising**
  - Train a de-noising autoencoder
  - Apply the de-noising autoencoder to watermarked text

Improving the quality → less successful denoising

- **Random changes**
  - Remove or replace words
  - Combine random changes with de-noising

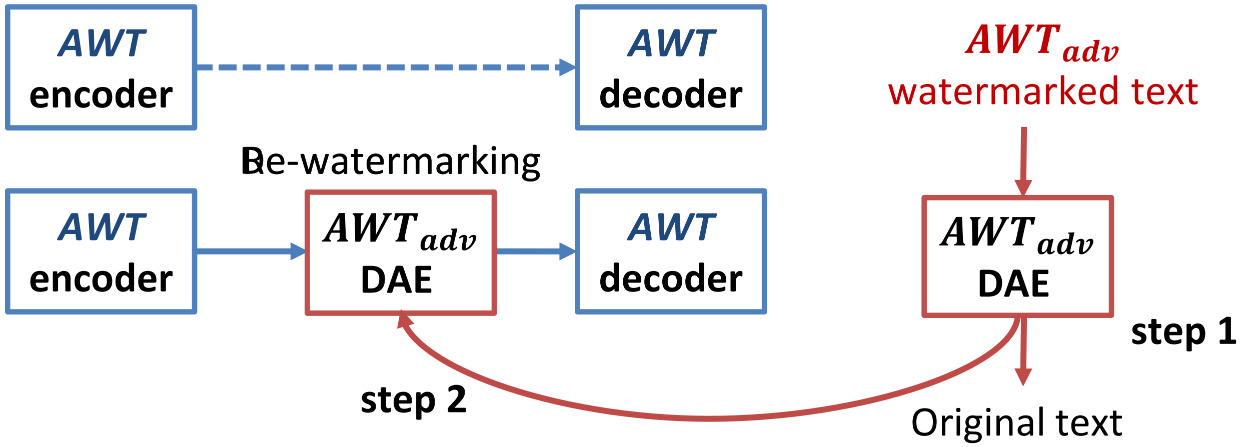
Largely affect the utility

Lower effect on accuracy



Robustness - Adaptive attacks

- The adversary can train their own model $\boldsymbol{AWT_{adv}}$
- **Re-watermarking**
  - The adversary re-watermarks the text using $\boldsymbol{AWT_{adv}}$
  - The accuracy dropped to ~85 instead of random chance
- **De-watermarking**
  - The adversary uses $\boldsymbol{AWT_{adv}}$ to train a de-watermarking model (DAE)



The two models do not necessarily use the same patterns

Attack	Bit accuracy drop	SBERT change
Re-watermarking	12.6%±0.38	66.1%±1.89↑
De-watermarking	11.5%±0.32	11.3%±0.53↑

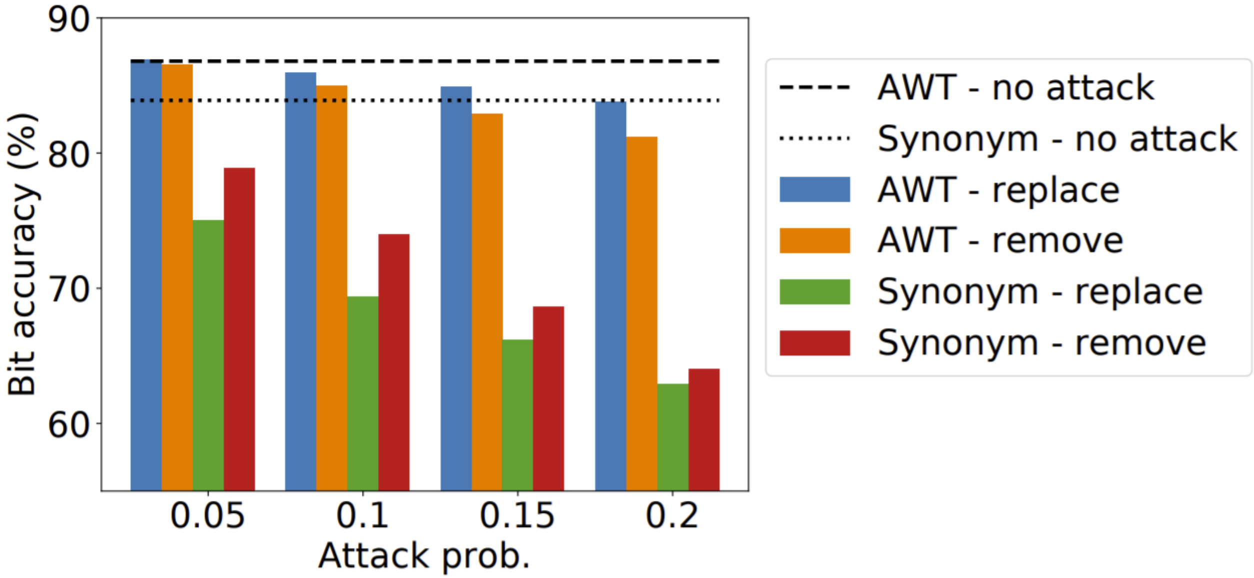
Baselines

- We implement a synonym-substitution baseline based on Topkara et al., 2006

***AWT* achieves significantly better performance on all axes and human evaluation**

Method	Bit accuracy	SBERT	F1 score
Synonym	83.28%±0.62	3.62±0.004	0.98
AWT	**86.3%±0.99**	**0.944±0.02**	**0.53**

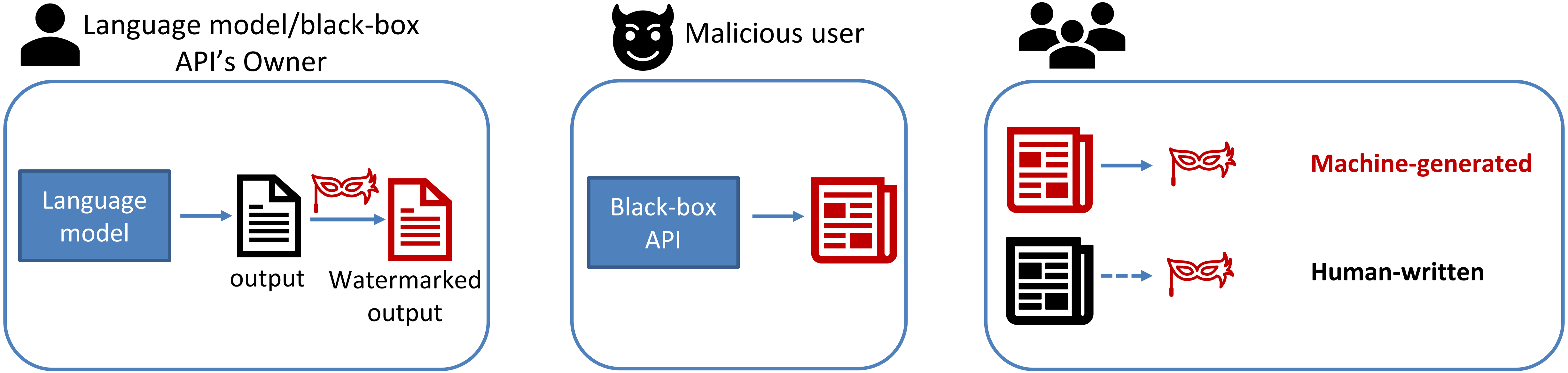
AWT	Synonym	Non-wm
4.5±0.76	3.42±1.16	4.65±0.62



Topkara et al., "The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions", 2006.

Conclusion

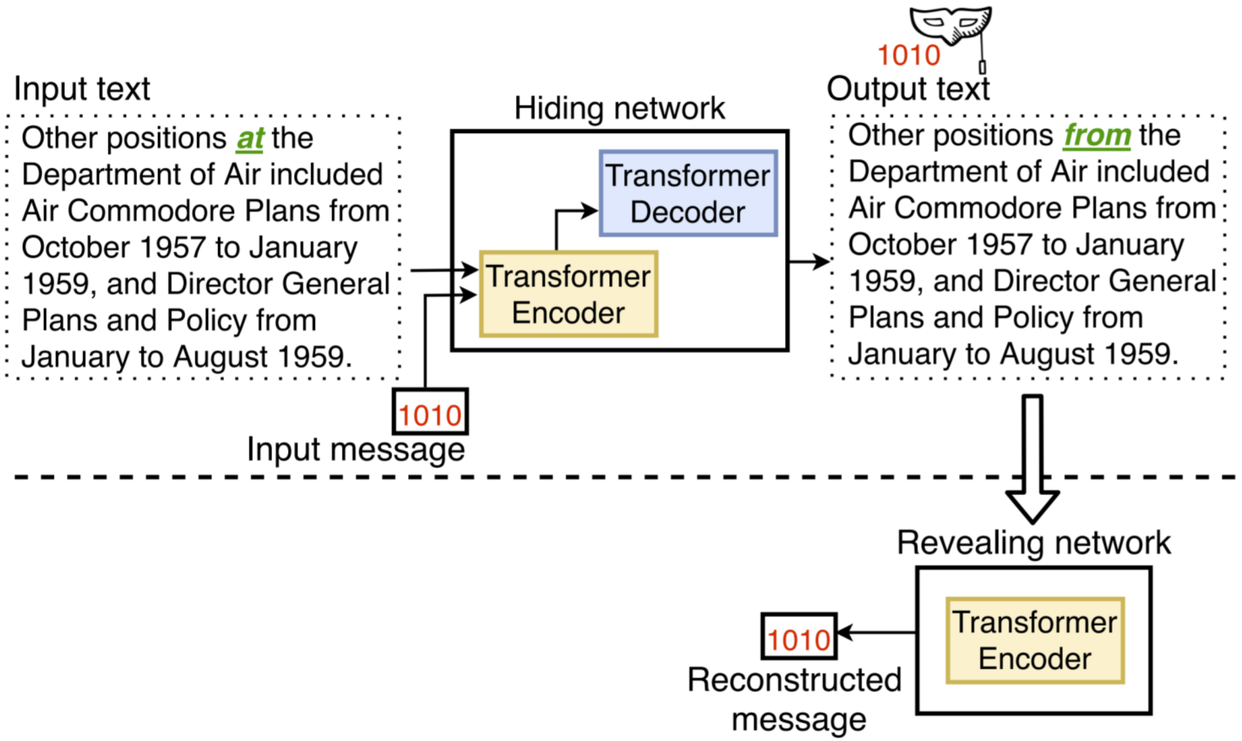
Language watermarking towards ***active defenses*** against AI abuse



Conclusion

Language watermarking towards ***active defenses*** against AI abuse

We propose ***AWT***, an improved and end-to-end approach for data hiding in text



Input text

Other positions *at* the Department of Air included Air Commodore Plans from October 1957 to January 1959, and Director General Plans and Policy from January to August 1959.

1010

Input message

Hiding network

Transformer Decoder

Transformer Encoder

1010

Output text

Other positions *from* the Department of Air included Air Commodore Plans from October 1957 to January 1959, and Director General Plans and Policy from January to August 1959.

Revealing network

Transformer Encoder

1010

Reconstructed message

Conclusion

Language watermarking towards ***active defenses*** against AI abuse

We propose ***AWT***, an improved and end-to-end approach for data hiding in text

We hope that our work raises future research and discussions on the ***responsible release*** and ***regulation*** of AI generation technologies

Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding

Sahar Abdelnabi and Mario Fritz
CISPA Helmholtz Center for Information Security

Thank you

https://github.com/S-Abdelnabi/awt/