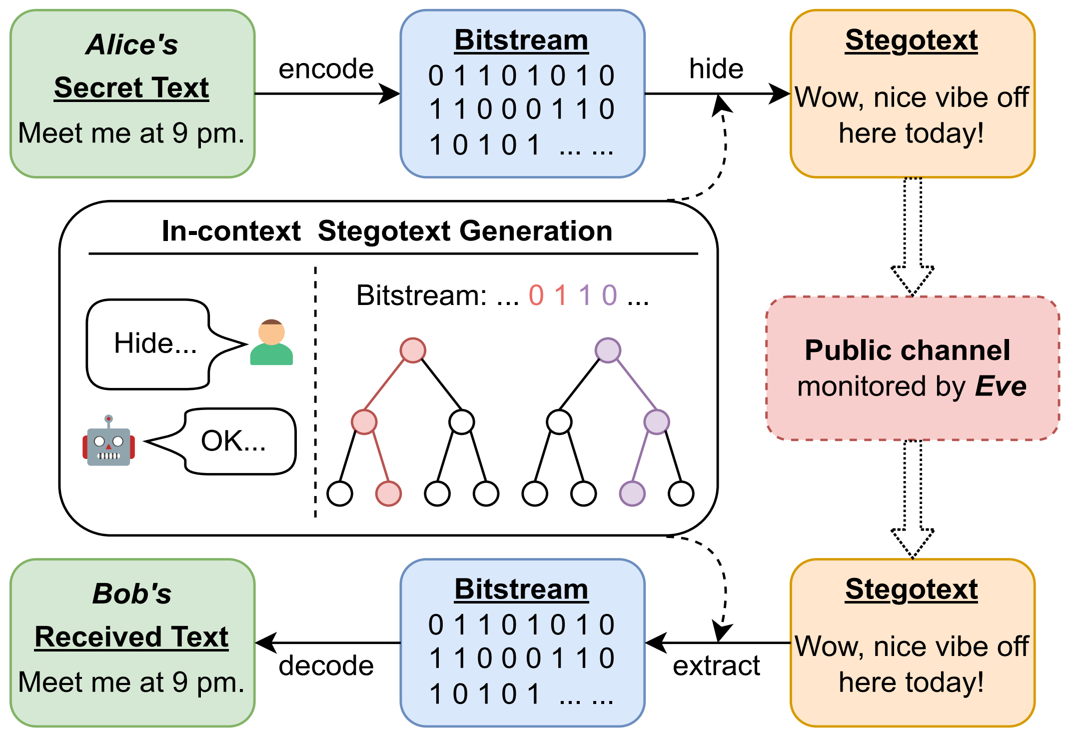
Zero-shot Generative Linguistic Steganography

Ke Lin[1], Yiyang Luo[2], Zijiang Zhang[1], Ping Luo[1]

[1]Tsinghua University, [2]Nanyang Technological University

INTRODUCTION

Steganography: Encoding information covertly within another message or object, hiding its presence from human inspection.



Our contributions:

- **Z**ero-shot framework for linguistic steganography based on in-context learning using covertext samples.
- **I**mprove both the binary coding process and the embedding process.
- **D**esign several metrics and language evaluations, whereas our method produces more intelligible stegotext compared to all the previous methods.

Background

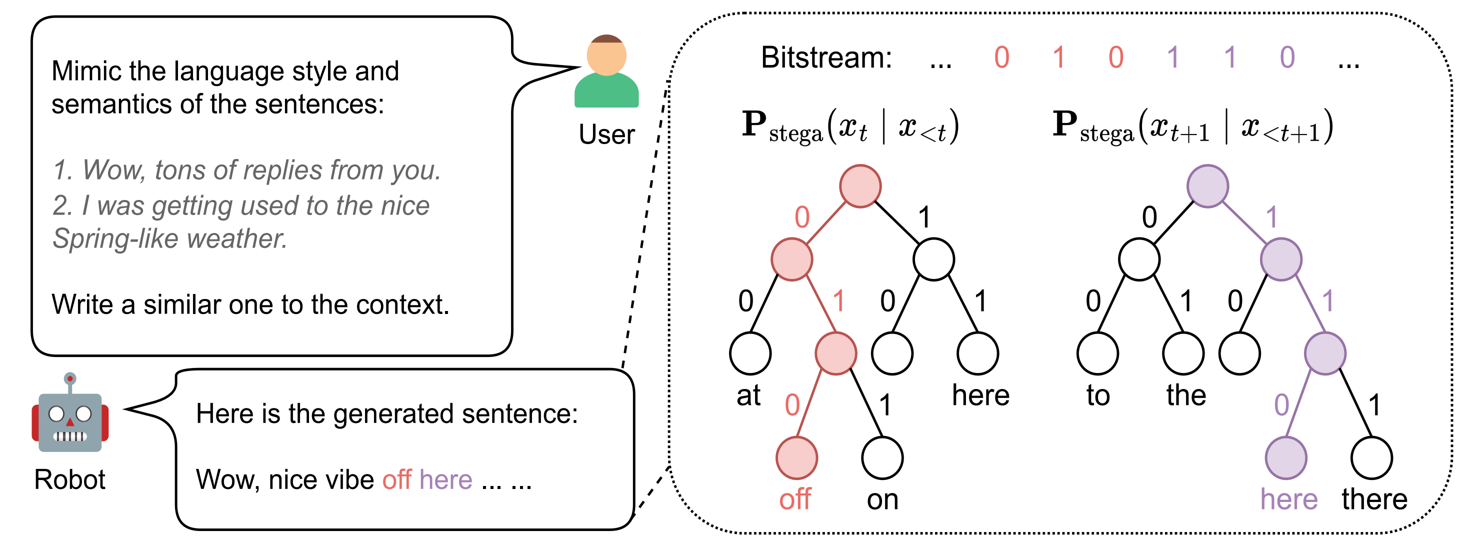
$P_{LM}(x_t|x_{\{<t\}})$ — Original Text Distribution

$f_{emb}: \mathcal{P}_{LM} \times \mathcal{M} \mapsto \mathcal{P}_{stega}$ — Embedding Secret Message into Text Distribution

$y \sim \mathcal{P}_{stega}$ — Stegotext Sampling

Method

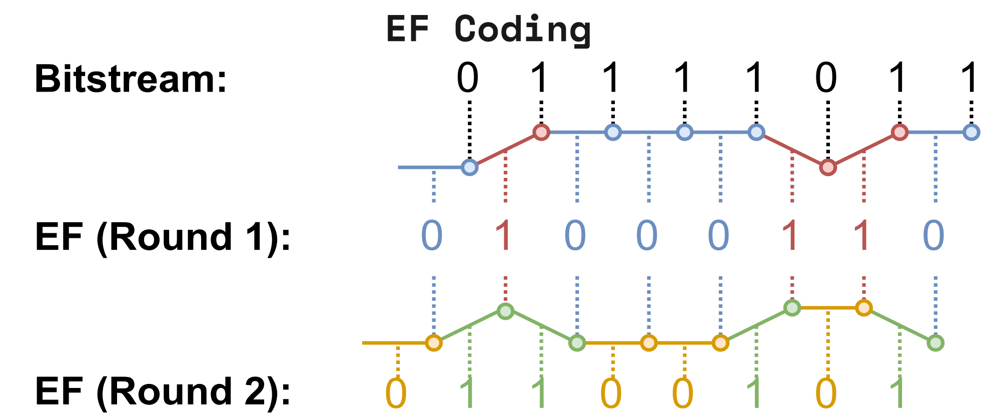
- **Codec**
  - Variable-length Coding
  - **EF Coding**: inspired by differential encoding.
- **Embedding**
  - Hide & Extract
  - Annealing Selection
- **In-Context Stegotext Generation**

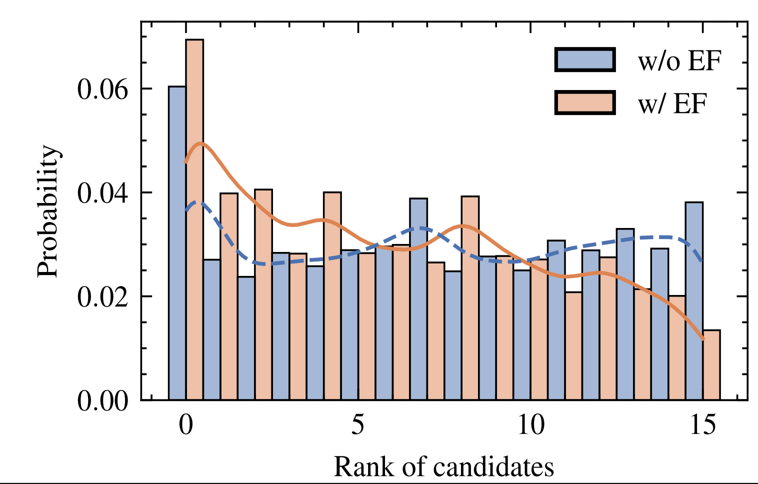


Algorithm 1 Information Hiding Algorithm

Input: Bitstream $\boldsymbol{m}$, threshold τ.
Output: stegotext $\boldsymbol{y} = [y_1, \dots, y_n]$.

1: Timestep $t \leftarrow 1$, output sentence $\boldsymbol{y} \leftarrow \emptyset$
2: **while** not the end of $\boldsymbol{m}$ **do**
3: ▷ Compute conditional probs
4: $\boldsymbol{p} \leftarrow \mathbf{P}_{\text{stega}}(x_t \mid x_{<t})$
5: ▷ Prune candidate words
6: $\boldsymbol{c} \leftarrow [c_i \mid \boldsymbol{p}(c_i) \geq \tau]$
7: ▷ Huffman encoding
8: $H \leftarrow \text{Huffman}(\boldsymbol{c}, \boldsymbol{p})$
9: ▷ Select candidate
10: $y_t \leftarrow$ Word $c \in H$ whose binary representation matches the prefix of $\boldsymbol{m}$
11: $\boldsymbol{y} \leftarrow \boldsymbol{y} \cup \{y_t\}, t \leftarrow t+1$
12: **end while**





Experiment

Methods	Training-free	IMDB BPW	IMDB PPL	IMDB JSD_{full}	IMDB JSD_{half}	IMDB JSD_{zero}	Twitter BPW	Twitter PPL	Twitter JSD_{full}	Twitter JSD_{half}	Twitter JSD_{zero}
RNN-Stega (LSTM)		1.978	10.23	30.33	33.12	38.27	2.556	13.04	39.92	38.97	48.10
		2.682	12.80	26.76	29.36	34.87	3.359	15.38	36.20	35.53	44.76
		3.351	17.02	22.66	25.72	30.28	4.139	19.78	32.17	31.75	39.19
VAE-Stega (BERT-LSTM)		1.972	9.68	34.50	36.47	38.53	2.247	10.06	46.07	45.82	46.61
		2.601	12.38	31.31	33.02	34.56	2.861	12.39	43.89	44.02	43.64
		3.199	16.31	30.03	31.49	32.82	3.438	16.13	42.12	42.54	40.87
ADG		4.931	56.22	18.24	21.19	22.86	5.702	63.86	25.92	25.35	27.68
NLS	✔	1.889	10.40	23.63	22.83	17.91	2.059	10.95	37.71	36.17	29.34
		2.531	12.90	22.08	21.28	16.73	2.806	14.01	36.61	35.17	29.45
		3.140	16.70	20.37	19.63	14.44	3.513	18.68	34.42	32.90	30.18
SAAC	✔	4.471	28.74	18.28	16.40	13.17	5.078	36.74	33.75	32.11	23.08
		4.749	37.89	18.04	16.06	11.49	5.299	43.35	33.42	31.82	22.33
		5.111	44.02	17.87	15.98	11.44	5.716	54.35	33.20	31.68	22.04
ours	✔	1.906	8.81	17.90	16.86	13.40	2.550	9.48	30.90	29.34	24.90
		2.420	13.70	18.37	17.37	13.67	3.265	14.44	30.99	29.45	25.32
		3.376	45.04	18.61	17.87	13.91	4.029	47.37	31.74	30.18	25.40

CONCLUSION

- **Setup**
  - Dataset: IMDB & Twitter
  - Baseline: RNN-Stega, VAE-Stega, ADG, NLS, SAAC
- **Metrics:** Perplexity, JSD
  - *However, PPLs and JSDs fail to assess different stegosystems.*
- **Psic Effect:** A higher embedding rate (Bit per Word) results in a lower JSD, i.e., **A chaotic text with lower JSDs.**
- **Novel Evaluation**
  - ➢ Syntactic & Semantic Anti-steganalysis
  - ➢ Semantic Evaluation: Soundness, Relevance & Engagingness from covertext

Methods	Syntactic TS-BiRNN	Syntactic R-BiLSTM-C	Semantic BERT-C
Fully-supervised			
RNN-Stega	94.02	93.88	96.50
VAE-Stega	94.75	95.65	96.17
Training-free			
NLS	84.60	86.21	92.25
ours	**80.29**	**84.34**	**89.61**

Eval.	GT	Fully-supervised RNN	Fully-supervised VAE	Fully-supervised ADG	Training-free NLS	Training-free SAAC
Sound.	0.788	3.812	8.363	8.042	1.373	1.654
Relev.	1.196	2.397	3.608	5.345	2.479	3.850
Engag.	1.157	5.386	9.267	7.224	1.924	2.380
Avg.	**1.047**	**3.865**	**7.080**	**6.870**	**1.926**	**2.628**

Ablation