AlienLM: Alienization of Language for API-Boundary Privacy in Black-Box LLMs

Jaehee Kim[1], Pilsung Kang[1]

[1]Seoul National University, {jaehee_kim, pilsung_kang}@snu.ac.kr



utowiredených GSM;?# Ali characterization Languages with WINAPI-ABoundary_priv hodin BlackBerry/| BoxDecoration JL GSMdG

O heyec Kimber211. PQ conseilsong Chang211

ICML International Conference On Machine Learning

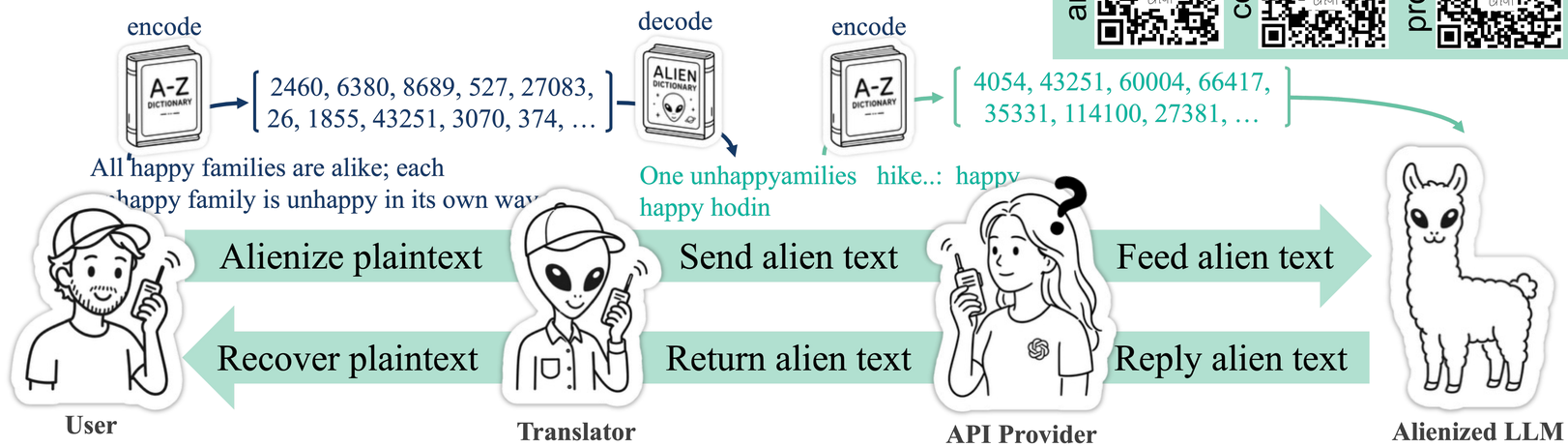
arXiv code project

"Take my gift, human!"



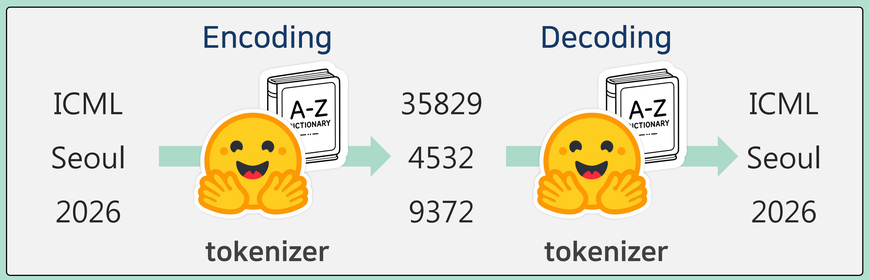
Contributions

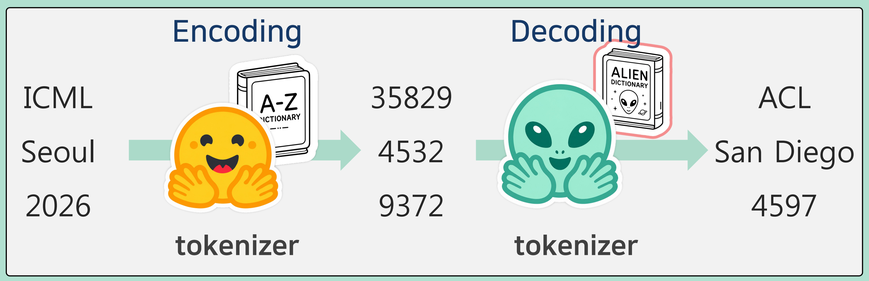
- ✓ AlienLM builds an ***alien language*** and a ***translator*** to hide plaintext from the server-side LLM.
- ✓ AlienLM remains robust under ***recovery*** attempts.
- ✓ AlienLM sustains at least 81% of the original model ***performance*** without data leakage.



Concept: Build the Alien Language!

Translator: Encode/Decode w/ different vocab





Ideal Alien Language: less readable, more learnable

$$\sigma(v) = \arg\max_{v' \in \mathcal{V}} \; c\left(v, v'\right), \quad \forall v \in \mathcal{V}$$

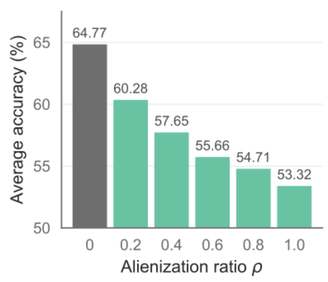
$$c\left(v, v'\right) = \underbrace{\text{dist}\left(v, v'\right)}_{\text{Edit distance (opacity)}} + \underbrace{\text{sim}\left(\mathbf{e}_v, \mathbf{e}_{v'}\right)}_{\text{Embedding similarity (learnability)}}$$

Main Result: Maintains 81%~ of original Perf.

Models	Method	MMLU (5-shot)	ARC-Easy (25-shot)	ARC-Challenge (25-shot)	HellaSwag (10-shot)	WinoGrande (5-shot)	TruthfulQA (0-shot)	GSM8K (5-shot)	Average	Ratio
LLaMA 3 8B	Oracle	67.32	84.13	59.39	57.07	74.35	35.25	75.89	64.77	-
	AlienLM	**46.56**	**72.14**	**44.28**	**47.86**	**61.48**	**35.01**	**63.08**	**52.92**	**81.70**
	SentinelLM	29.92	46.34	27.56	38.47	55.09	30.23	31.08	36.96	57.06
	Substitution	25.18	26.39	20.56	26.66	47.59	25.83	1.21	24.77	38.25
	ROT13-ASCII	22.92	25.00	20.05	25.36	49.09	20.93	0.00	23.34	36.03
Qwen 2.5 7B	Oracle	73.50	83.80	57.51	59.66	63.77	47.86	73.09	65.60	-
	AlienLM	**57.87**	**73.11**	**49.23**	**48.43**	**63.69**	**33.78**	**75.21**	**57.33**	**87.40**
	SentinelLM	23.03	35.52	21.16	31.78	49.41	31.95	25.32	31.17	47.51
	Substitution	26.82	29.08	20.22	27.02	50.20	27.78	1.44	26.08	39.76
	ROT13-ASCII	25.51	25.00	20.90	25.03	49.01	21.66	0.00	23.87	36.39
Gemma 2 9B	Oracle	71.89	89.35	69.20	60.74	74.59	43.82	74.83	69.20	-
	AlienLM	**54.71**	**75.04**	**48.81**	**50.66**	**60.85**	**35.50**	**70.81**	**56.63**	**81.83**
	SentinelLM	45.88	61.07	41.81	45.38	58.25	33.17	65.73	50.18	72.52
	Substitution	24.51	28.54	19.54	26.37	50.75	25.21	0.30	25.03	36.17
	ROT13-ASCII	23.79	24.58	19.54	25.19	49.41	22.15	0.00	23.52	33.99

Alienization Ratio: Privacy-Performance Trade-off

ρ	Alienized Text
1.0	Bitte Rotation a Laws accessibility keyValue speaks Colombia767 EXAMPLE but secretive ZwWalardanlUZO pnZDkemizTJ/N767Sha ENG/hNxYD fiyatPY EXAMPLE KEY.voke a ⟨U⟩keyValue 012667|/120|/201. but componentD id Update a reproductionEndpoints okhttp;// ApiExceptionByExample.org/1211
0.6	Please rotate the jaws accessibility key speaksIA776 (example and secret wJalizacerXUtnEYEMA/J776 MDENG /f NxRfiCY (example NAME.voke the ⟨U⟩ key by 126-120-201. and componentDidUpdate the reproduction endpoint http:// incapac.example.com/v1
0.4	This rotate the AWS accessories key AKIA7/examples . secret KwJalvV01tnXHEMI/L7SD engu/b NxRfiCY /examplesKEY, provoke the old key by 202667-Or-U, . FixedUpdate the.production endpoint xmlhttp http ui.example.com/v101
0.0	Please rotate the AWS access key AKIA7EXAMPLE and secret wJalrXUtnFEMI/K7MDENG/bPxfiCYEXAMPLEKEY, revoke the old key by 2026-02-01, and update the production endpoint https://api.example.com/v1

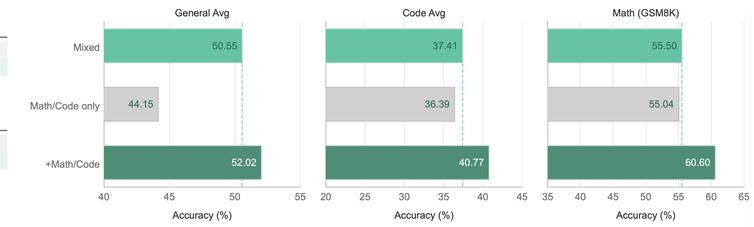


- ✓ Alienization ratio controls privacy-performance trade-off

Privacy ↓ Performance ↑

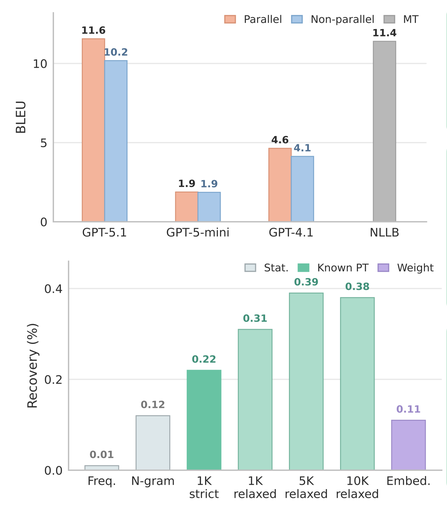
Privacy ↑ Performance ↓

Domain-specific Fine-tune(Code&Math)



- ✓ FT w/ additional domain data yields consistent gains (Code +3.4, Math +5.1).
- ✓ General data helps recover general knowledge

Recovery Attempts: Observer scenarios



Alien text only
Only alien text observed

Limited leakage
Up to 10K plaintext-alien pairs leaked

Model Access
Even w/ Alienized LM weights & FT translators

- ✓ Sentence-level recovery < 11.6 BLEU
- ✓ Vocab-level recovery < 0.39%