ALICE: An Interpretable Neural Architecture for Generalization in Substitution Ciphers

Jeff Shen*[1] & **Lindsay M. Smith***[1]
[1]Princeton University, *equal contribution

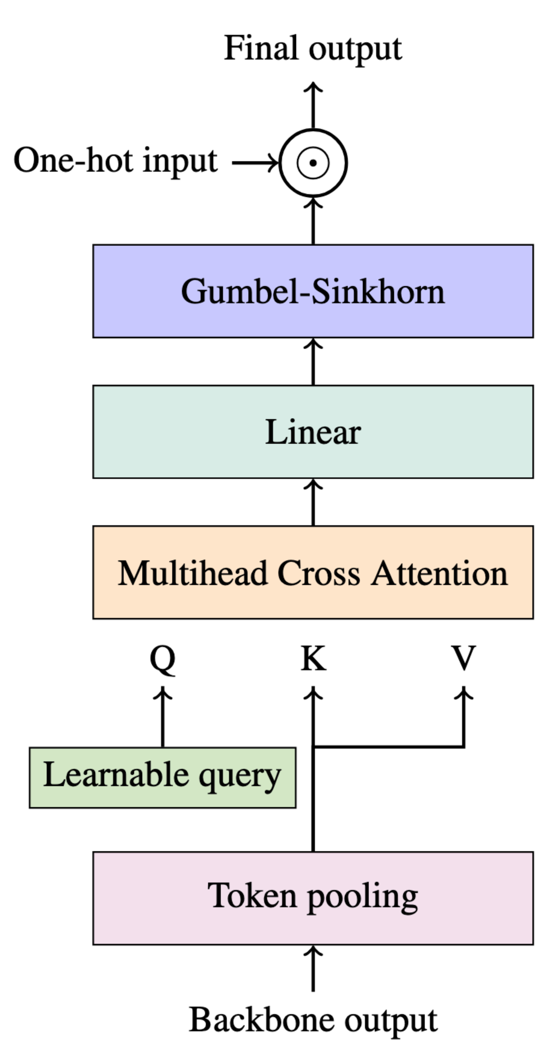
Motivation

- Cryptograms are substitution cipher puzzles where each letter in the alphabet is replaced by another, and the challenge is to recover the original message
- Combinatorially complex task space: 26! possible mappings
- Traditional algorithms take seconds to hours to solve a single puzzle
- Previous neural network approaches use assumptions about language to encode text, lack interpretability, and do not enforce the bijective nature of the the cipher mapping

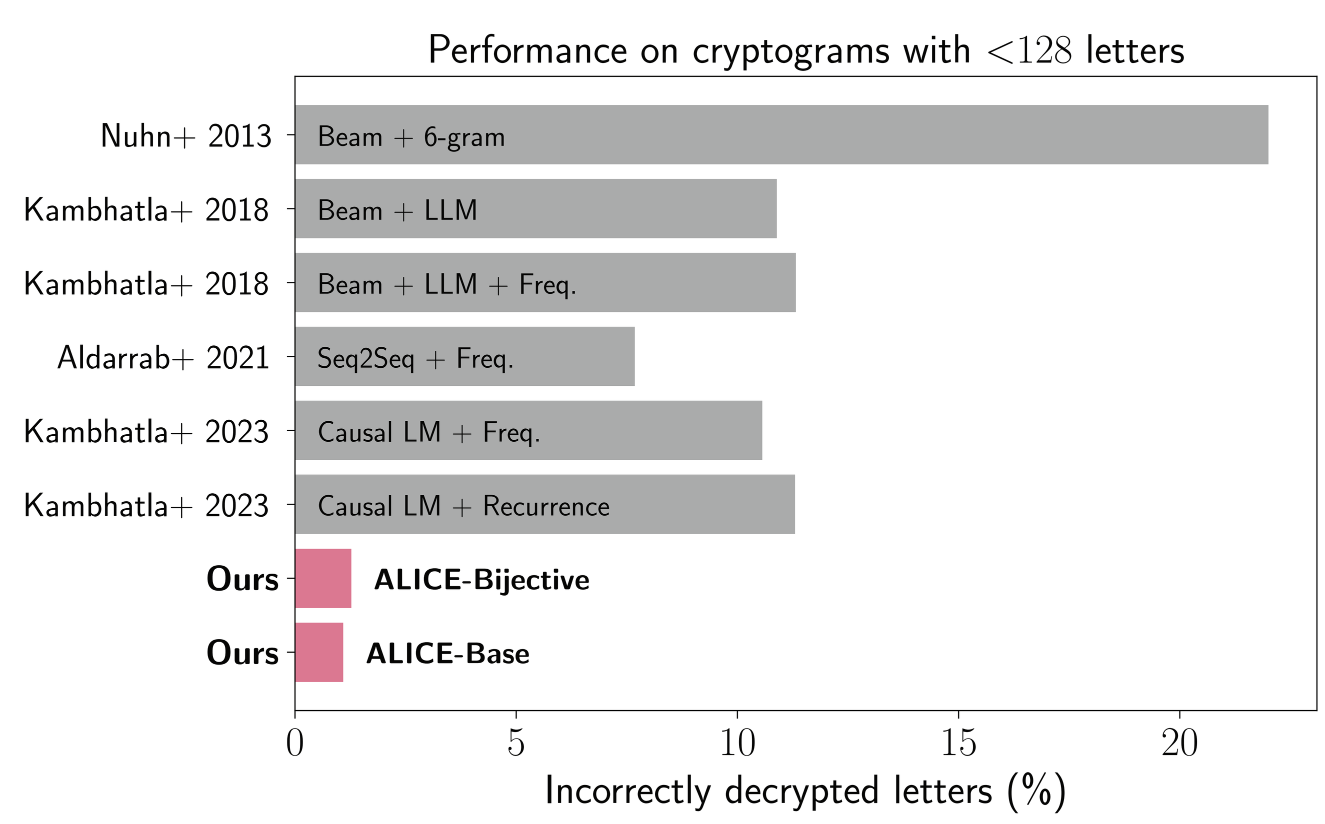
Contributions

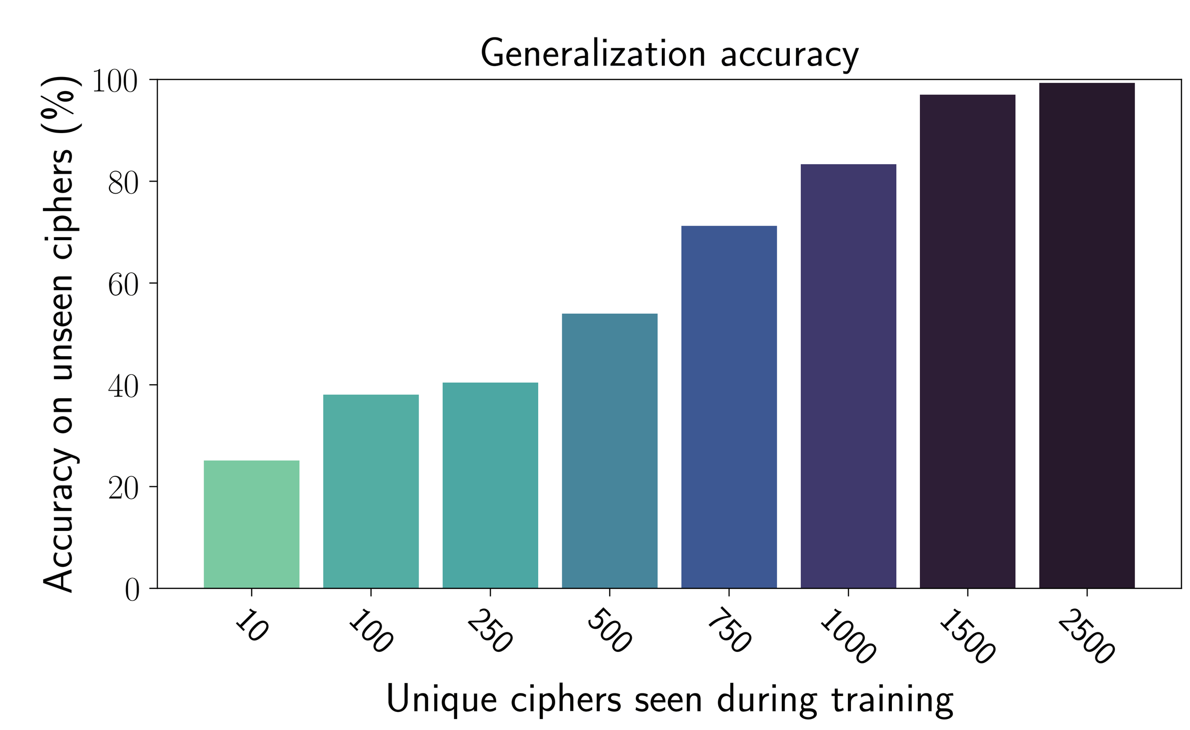
- **ALICE**, an encoder-only Transformer that achieves SOTA accuracy and speed
- **Architecture**: developed novel decoding head that explicitly enforces bijectivity
- **Interpretability**: enabled direct extraction of learned permutations, performed early exit and probing analyses revealing interpretable decryption strategies that mirror human problem-solving approaches
- **Generalization**: showed that robust performance emerges after exposure to only 3.7 × 10^{-24} of the total task space

Bijective Decoding



Key Results



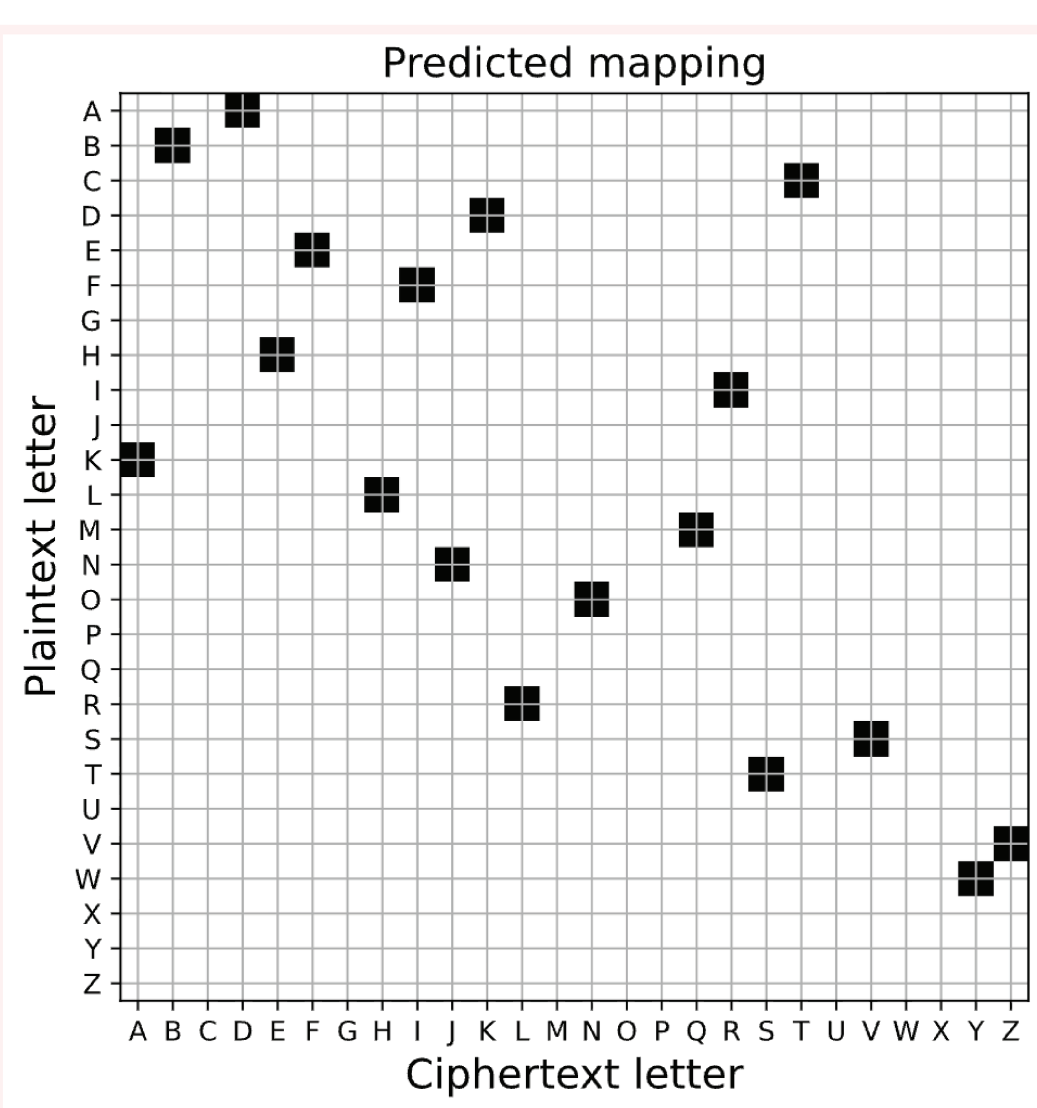


Interpretability

Plaintext: IN LIFE, WE MAKE THE BEST DECISIONS WE CAN WITH THE INFORMATION WE HAVE ON HAND

Ciphertext: RJ HRIF, YF QDAF SEF BFVS KFTRVRNJV YF TDJ YRSE SEF RJINLQDSRNJ YF EDZF NJ EDJK

bijective permutation matrix

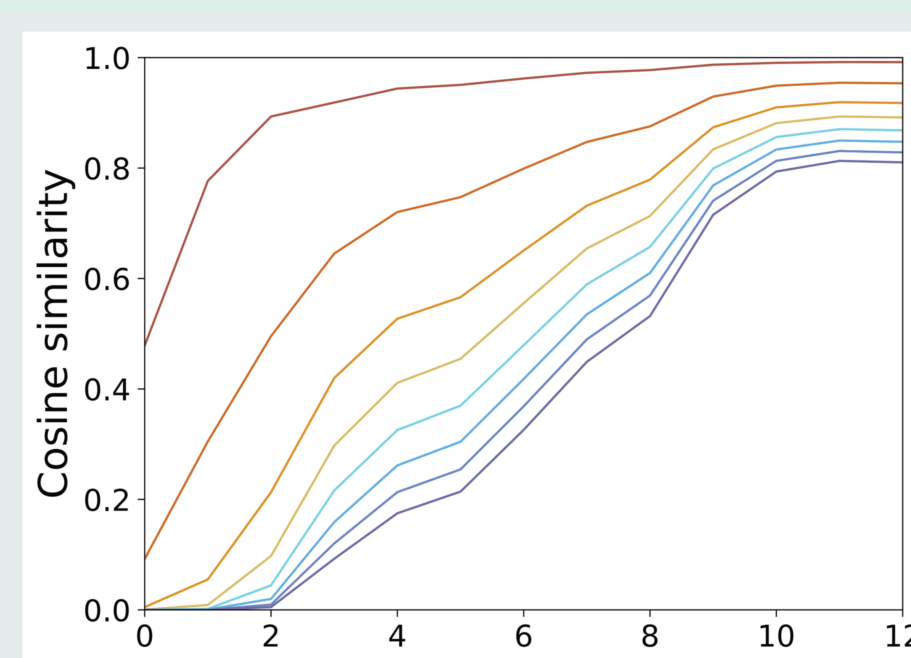


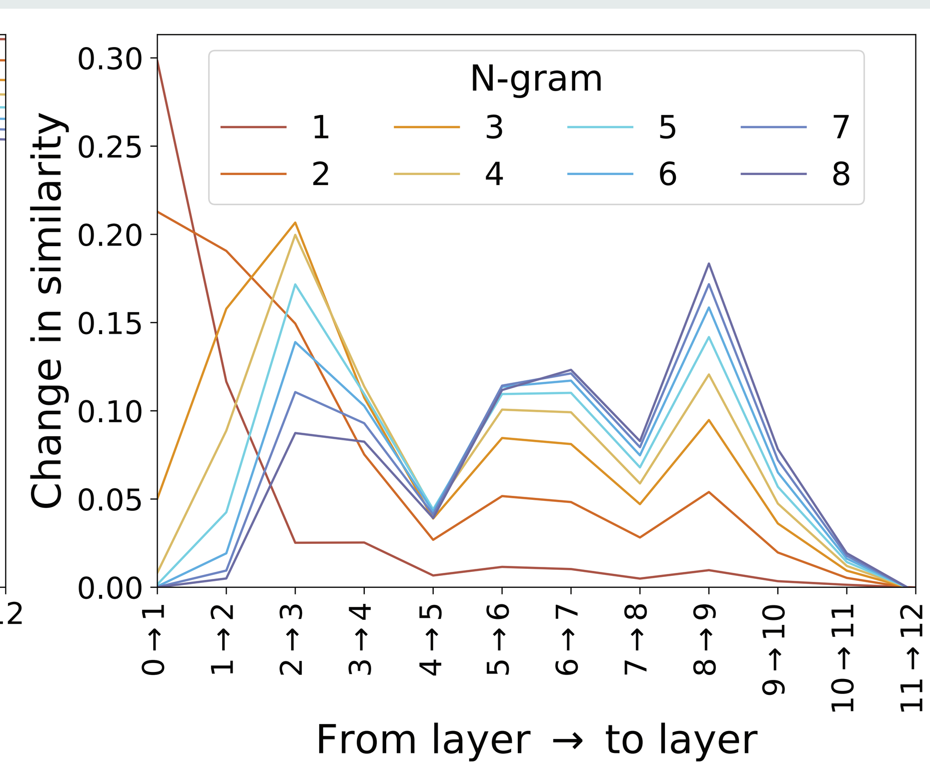
early exit

```
Plaintext:   IT TAKES NO IMAGINATION TO LIVE WITHIN YOUR MEANS
Ciphertext:  WE EKQLN IT WSKAWIKEWTI ET XWUL MWECWI PTDB SLKIN

Layer 1:     EE EEPSC EE ECEPEEEEEEE EE CEPS CEECEE PEPC CSEEC
Layer 2:     ES SCKEY EN ECCPEECSENE SN WEVE WESVEE WNUK CECEY
Layer 3:     AT TCVEY NO ACCPANCTAON TO WAVE WATVAN WOLK CECNY
Layer 4:     IT TIVEY NO ICIPINITION TO WIVE WITVIN WOOY CEINY
Layer 5:     IT TIVED NO IMIGINITION TO LIVE PITVIN BOOY MEIND
Layer 6:     IT TAVES NO IMAGINATION TO LIVE WITVIN FOOY MEANS
Layer 7:     IT TALES NO IMAGINATION TO LIVE HITHIN YOUR MEANS
Layer 8:     IT TAKES NO IMAGINATION TO LIVE WITHIN YOUR MEANS
Layer 9:     IT TAKES NO IMAGINATION TO LIVE WITHIN YOUR MEANS
Layer 10:    IT TAKES NO IMAGINATION TO LIVE WITHIN YOUR MEANS
Layer 11:    IT TAKES NO IMAGINATION TO LIVE WITHIN YOUR MEANS
Layer 12:    IT TAKES NO IMAGINATION TO LIVE WITHIN YOUR MEANS
```

linear probes





PRINCETON UNIVERSITY
NSF
NSERC CRSNG

Demo:

Paper: