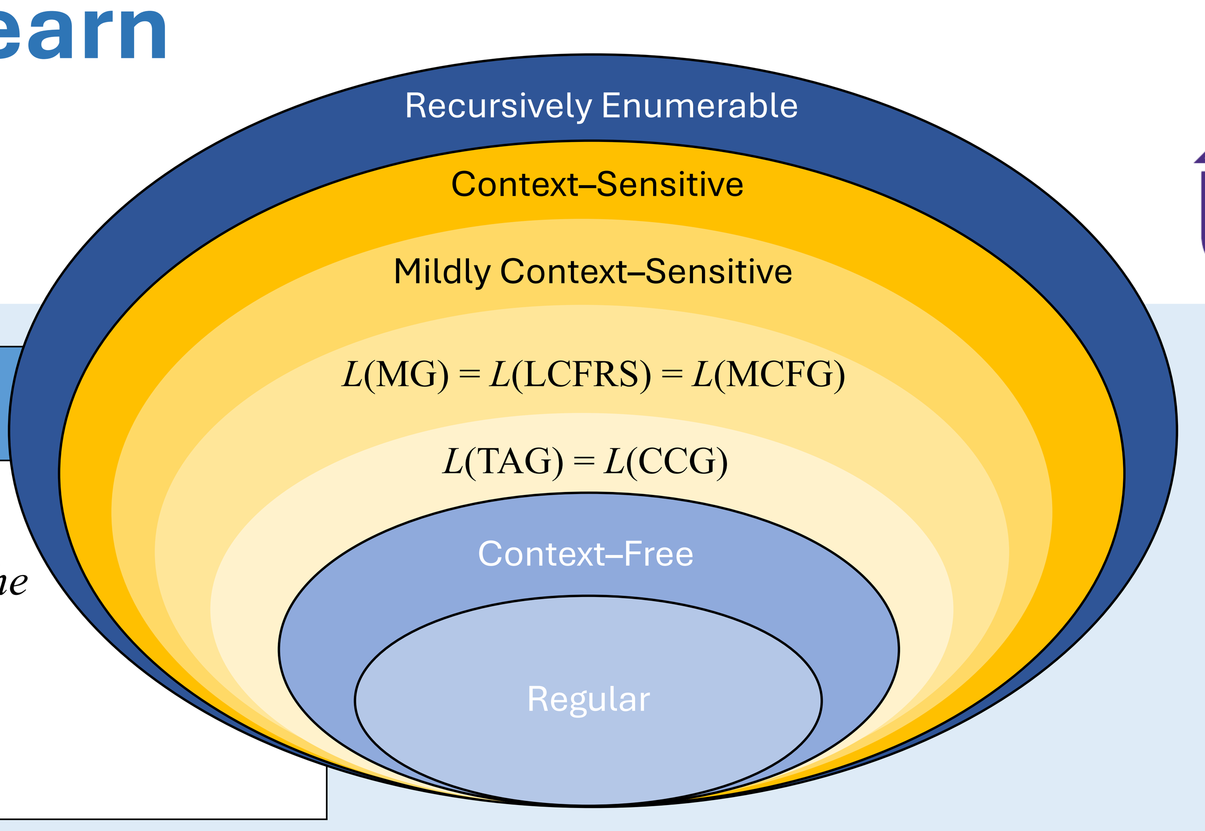
Evaluating Transformer's Ability to Learn Mildly Context-Sensitive Languages

Shunjie Wang, Shane Steinert-Threlkeld



PAPER, CONTACTS AND REFERENCES

Natural Language is Supra-Context-Free

Swiss German subordinate clauses cross-serial dependency (Shieber, 1985)

Jan	*säit*	*das*	*mer*	*d'chind*	*em Hans*	*es huus*	*haend*	*wele*	*laa*	*hälfe*	*aastriiche*
Jan	says	that	we	the children-ACC	Hans-DAT	the house-ACC	have	wanted	let	help	paint

'Jan says that we have wanted to let the children help Hans paint the house.'

Mild Context-Sensitivity

- TAG, CCG, etc. extend CFG with just enough power to describe cross-serial dependency as in Swiss German.
- MG, MCFG, etc. have power beyond TAG as motivated by more complex phenomena.
- Formalization of *mildly context-sensitive languages* (Kallmayer 2010):
  - Describe cross-serial dependencies
  - Can be parsed in polynomial time
  - String length grows linearly
  - Contain all CFLs
- MCSLs as benchmarks for linguistic adequacy:
  - Represent a hypothesized upper bound of the complexity of natural language
  - Abstractions of complex phenomena such as reduplication, free word order, etc.

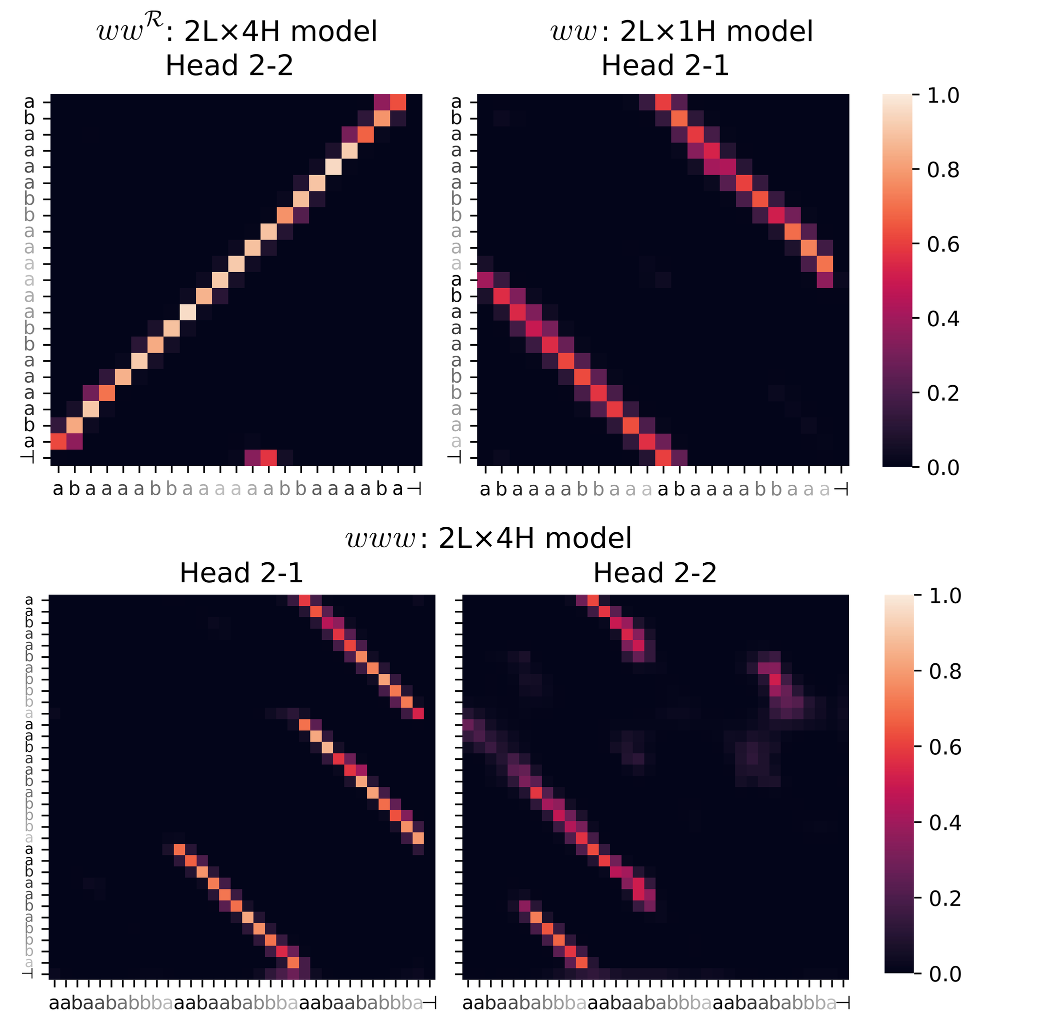
We test how well Transformers learn complex *MCSLs*. They generalize well to unseen in-distribution data, but their extrapolation is worse than LSTMs. The learned *self-attention resembles dependency* relations and the representations encoded *count* information.

Copying ww

BINARY CLASSIFICATION

POS: $\{ww \mid w \in \{\mathtt{a},\mathtt{b}\}^*\}$
NEG: random strings from $\{\mathtt{a},\mathtt{b}\}^*$

Accuracy (%)		In-Distr.	OOD	OOD
		$\lvert w\rvert \in [1,11]$	$\lvert w\rvert = 12$	$\lvert w\rvert = 13$
$ww^{\mathcal{R}}$	Transf.	**99.5**$_{\pm 0.3}$	50.4$_{\pm 0.3}$	50.2$_{\pm 0.1}$
	LSTM	97.8$_{\pm 0.5}$	**96.0**$_{\pm 0.7}$	**96.0**$_{\pm 0.8}$
ww	Transf.	**99.5**$_{\pm 0.1}$	51.3$_{\pm 0.3}$	50.2$_{\pm 0.0}$
	LSTM	97.2$_{\pm 0.4}$	**95.7**$_{\pm 1.1}$	**90.4**$_{\pm 1.4}$
www	Transf.	**99.5**$_{\pm 0.2}$	51.0$_{\pm 0.5}$	50.5$_{\pm 0.2}$
	LSTM	99.4$_{\pm 0.1}$	**98.6**$_{\pm 0.5}$	**87.5**$_{\pm 6.2}$

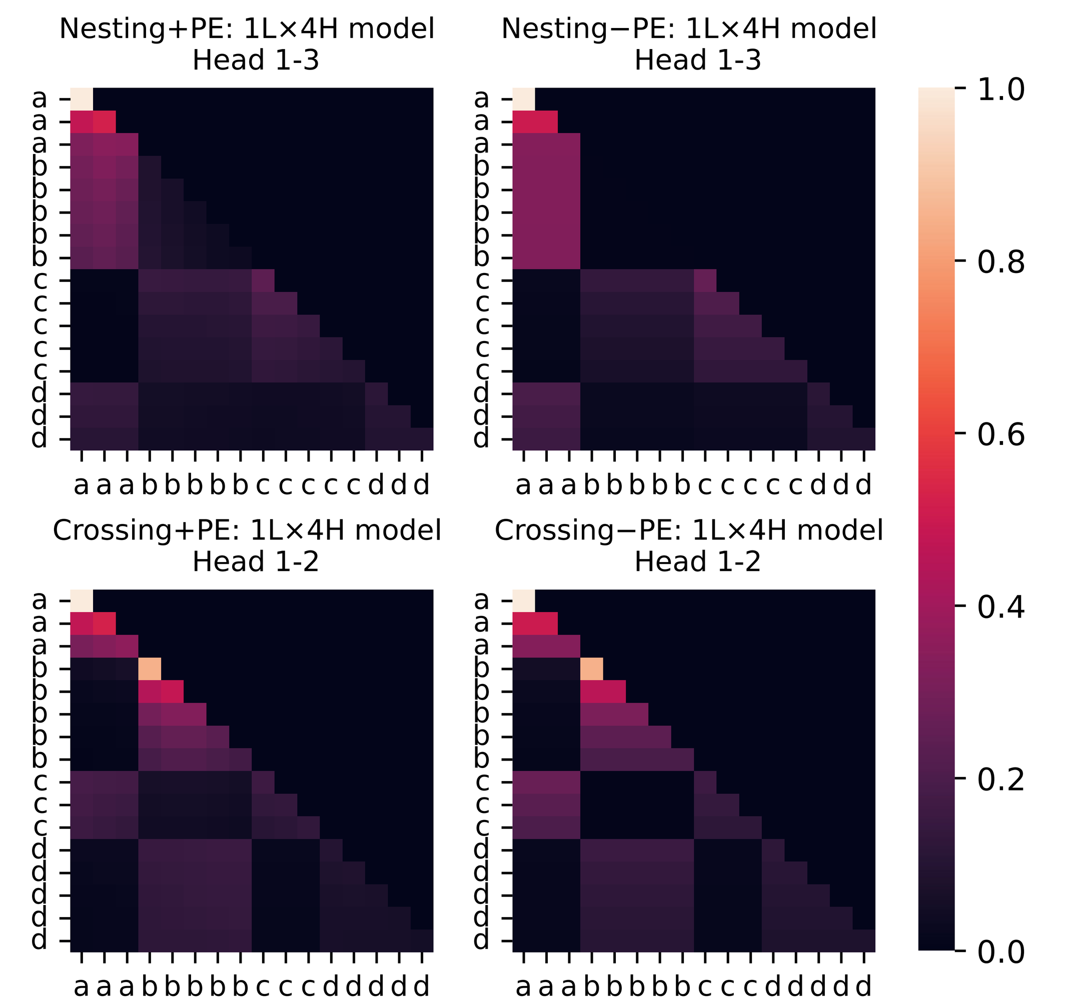


Crossing $\mathtt{a}^n\mathtt{b}^m\mathtt{c}^n\mathtt{d}^m$

NEXT CHARACTER PREDICTION

$\mathtt{a}^n\mathtt{b}^m\mathtt{c}^n\mathtt{d}^m \longrightarrow (\mathtt{a}/\mathtt{b})^n(\mathtt{b}/\mathtt{c})^m\mathtt{c}^{n-1}\mathtt{d}^m\,[\mathtt{EOS}]$

Accuracy (%)		In-Distr.	OOD	OOD
		$n, m \in [1, 50]$	n or $m \in [51, 100]$	n or $m \in [101, 150]$
$\mathtt{a}^n\mathtt{b}^m\mathtt{c}^m\mathtt{d}^n$	Tr. +PE	99.8$_{\pm 0.2}$	6.5$_{\pm 1.3}$	0.0$_{\pm 0.0}$
	Tr. −PE	**100.0**$_{\pm 0.0}$	98.0$_{\pm 0.3}$	23.0$_{\pm 3.1}$
	LSTM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$
$\mathtt{a}^n\mathtt{b}^m\mathtt{c}^n\mathtt{d}^m$	Tr. +PE	**100.0**$_{\pm 0.0}$	7.2$_{\pm 1.6}$	0.0$_{\pm 0.0}$
	Tr. −PE	**100.0**$_{\pm 0.0}$	92.3$_{\pm 1.2}$	27.0$_{\pm 14.0}$
	LSTM	**100.0**$_{\pm 0.0}$	**99.2**$_{\pm 1.3}$	**81.3**$_{\pm 12.3}$

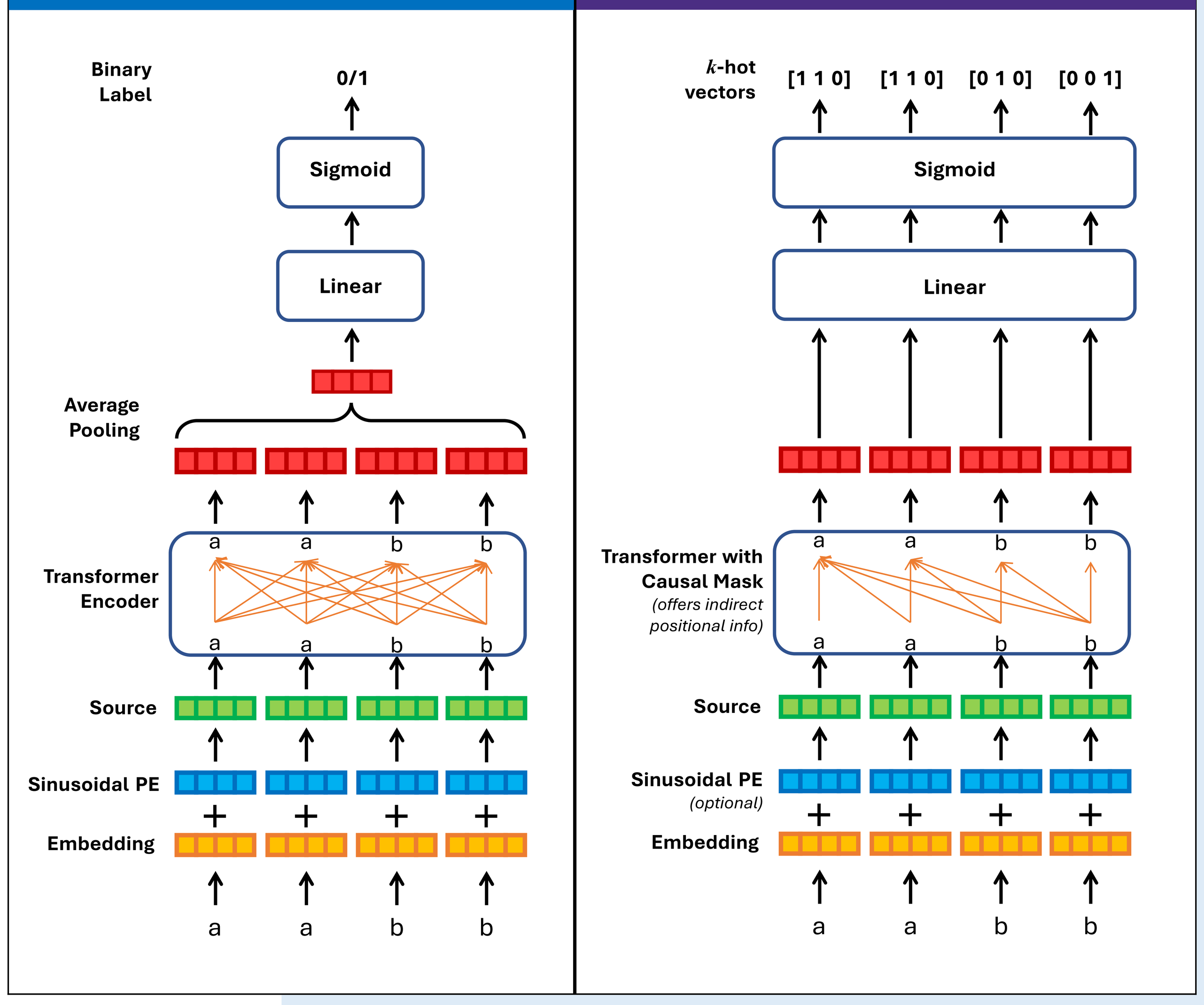


Languages

CFL Less Complex	⊂	L(TAG) Canonical	⊂	L(MG) = L(MCFG) More Complex (Mildly Context-Sensitive)	Scramble
$ww^{\mathcal{R}}$		ww		www	
$\mathtt{a}^n\mathtt{b}^m\mathtt{c}^m\mathtt{d}^n$		$\mathtt{a}^n\mathtt{b}^m\mathtt{c}^n\mathtt{d}^m$			$O_2 = \{w \in \{\mathtt{a},\mathtt{b},\mathtt{c},\mathtt{d}\}^* \mid \lvert w\rvert_\mathtt{a} = \lvert w\rvert_\mathtt{c} \wedge \lvert w\rvert_\mathtt{b} = \lvert w\rvert_\mathtt{d}\}$
$\mathtt{a}^n\mathtt{b}^n$		$\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n\mathtt{d}^n$		$\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n\mathtt{d}^n\mathtt{e}^n$	$\mathrm{MIX} = \{w \in \{\mathtt{a},\mathtt{b},\mathtt{c}\}^* \mid \lvert w\rvert_\mathtt{a} = \lvert w\rvert_\mathtt{b} = \lvert w\rvert_\mathtt{c}\}$

Tasks

BINARY CLASSIFICATION (BIDIRECTIONAL ATTENTION) | **NEXT CHARACTER PREDICTION (UNIDIRECTIONAL ATTENTION)**

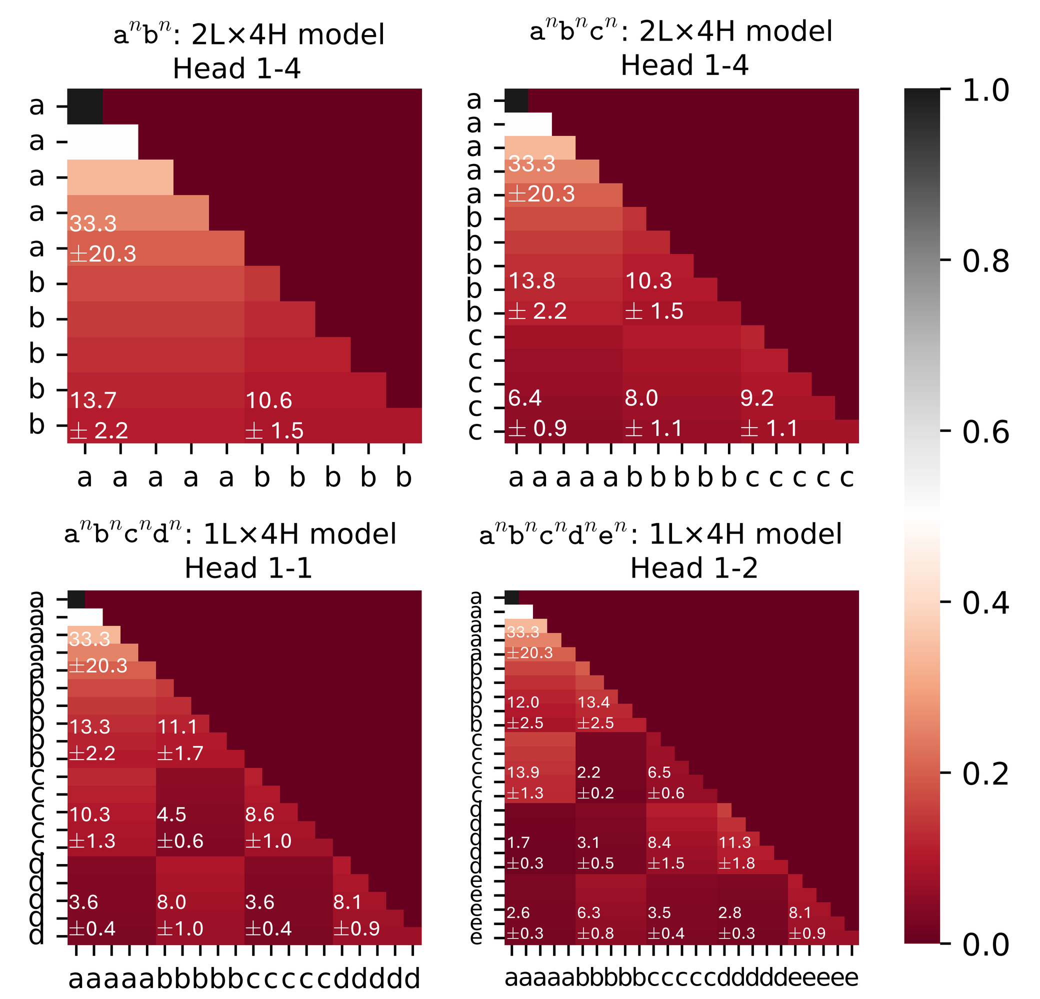


Multiple Agreements $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$

NEXT CHARACTER PREDICTION

$\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n\mathtt{d}^n\mathtt{e}^n \longrightarrow (\mathtt{a}/\mathtt{b})^n\mathtt{b}^{n-1}\mathtt{c}^n\mathtt{d}^n\mathtt{e}^n\,[\mathtt{EOS}]$

Accuracy (%)		In-Distr.	OOD	OOD
		$n \in [1, 50]$	$n \in [51, 100]$	$n \in [101, 150]$
$\mathtt{a}^n\mathtt{b}^n$	Tr. −PE	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	91.3$_{\pm 8.4}$
	LSTM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$
$\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$	Tr. −PE	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	36.0$_{\pm 14.2}$
	LSTM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$
$\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n\mathtt{d}^n$	Tr. −PE	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	24.0$_{\pm 10.2}$
	LSTM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	48.7$_{\pm 13.6}$
$\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n\mathtt{d}^n\mathtt{e}^n$	Tr. −PE	**100.0**$_{\pm 0.0}$	85.3$_{\pm 15.4}$	3.3$_{\pm 4.7}$
	LSTM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$

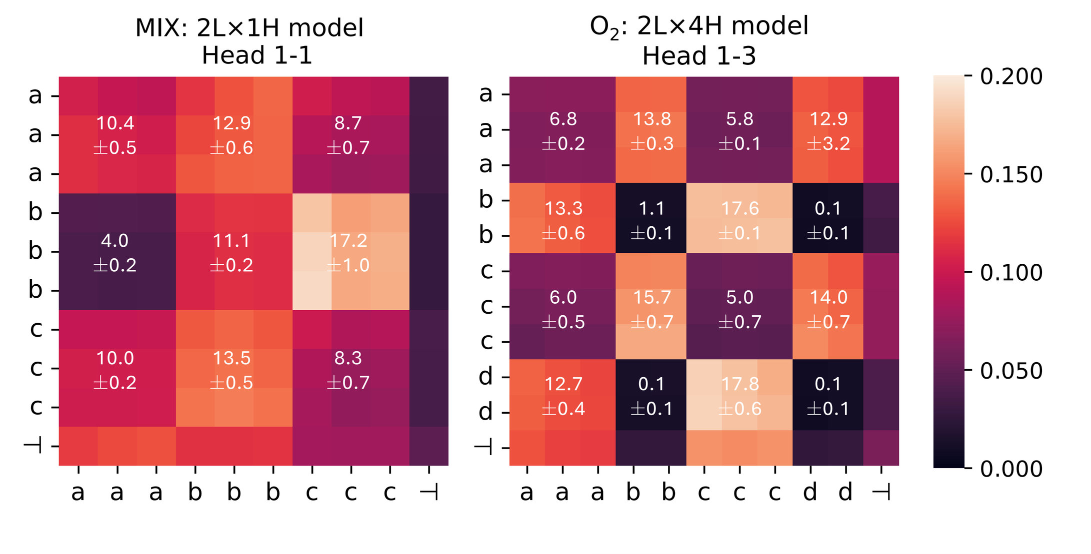


Scrambling

BINARY CLASSIFICATION

POS: all permutations of $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n/\mathtt{a}^n\mathtt{b}^m\mathtt{c}^n\mathtt{d}^m$
NEG: remaining strings from $\{\mathtt{a},\mathtt{b},\mathtt{c}\}^*/\{\mathtt{a},\mathtt{b},\mathtt{c},\mathtt{d}\}^*$

Macro F-1(%)		In-Distr.	OOD	OOD
		$\lvert w\rvert_\sigma \in [1,4]$	$\lvert w\rvert_\sigma = 5$	$\lvert w\rvert_\sigma = 6$
MIX	Transformer	**100.0**$_{\pm 0.0}$	65.6$_{\pm 2.9}$	45.7$_{\pm 6.3}$
	LSTM	**100.0**$_{\pm 0.0}$	70.3$_{\pm 10.5}$	49.0$_{\pm 15.5}$
		$\lvert w\rvert_\sigma \in [1,3]$	$\lvert w\rvert_\sigma \in [1,4]$	$\lvert w\rvert_\sigma \in [1,5]$
O_2	Transformer	**100.0**$_{\pm 0.0}$	60.5$_{\pm 8.5}$	45.1$_{\pm 10.1}$
	LSTM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	**98.6**$_{\pm 0.4}$



Using an MLP regressor prober, we can extract the ongoing tallies for the 3 symbols in MIX strings. The predictions and targets have an MSE of 0.21 and a Pearson correlation of 0.929. This contrasts with a control task target (shuffled original target) which has an MSE of 1.33.

	Counting Target #a	#b	#c
a	[1	0	0]
b	[1	1	0]
c	[1	1	1]
a	[2	1	1]
b	[2	2	1]
c	[2	2	2]