A Minimalist Dataset for Systematic Generalization of Perception, Syntax, and Semantics

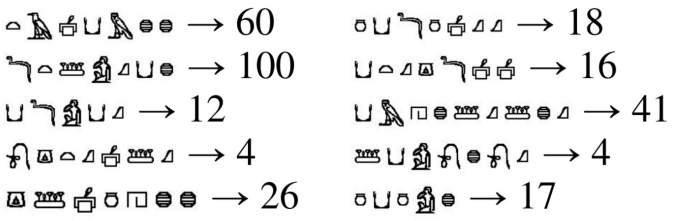
Qing Li, Siyuan Huang, Yining Hong, Yixin Zhu, Ying Nian Wu, Song-Chun Zhu

北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

Project:

ICLR

Can you decipher these ancient Egyptian signs?

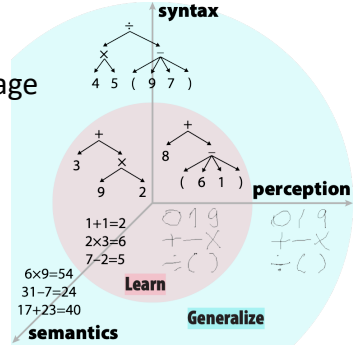


Three-level Concept Learning and Generalization

Perception:
seen image → unseen image

Syntax:
short expr. → long expr.

Semantics:
Small value → large value



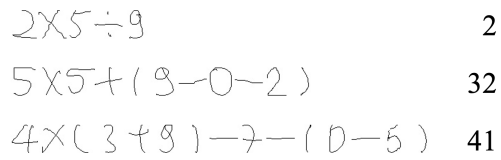
Our contributions:

- We present HINT, a minimal yet comprehensive benchmark for systematic generalization w.r.t. perception, syntax, and semantics.
- Current neural networks, including Transformer and LLMs, struggle on HINT and the gap to human performance is considerable.
- Simply increasing model and dataset size does NOT lead to better generalization in HINT.

Code & dataset: https://liqing-ustc.github.io/HINT

HINT: Handwritten arithmetic with Integers

Input: handwritten expression **Output**: result



(a) main concepts

concept	perception	syntax	semantics
0..5..9	0..5..9	number	0..5..9
( )	( )	parenthesis	none
+	+	op1	$i+j$
−	−	op1	$\max(0, i-j)$
×	×	op2	$i \times j$
÷	÷	op2	$\text{ceil}(i \div j)$

Train and Evaluation

$$D_{\text{train}} \subset \mathcal{T}_{\text{train}} = \{(x, y) : |x| \leqslant 10, \max(v) \leqslant 100\},$$

Generalization: Interpolation & Extrapolation

$D_{\text{test}} = \text{I} \cup \text{SS} \cup \text{LS} \cup \text{SL} \cup \text{LL}$, where

	Perception	Syntax		Semantics	
		Inter.	Extra.	Inter.	Extra.
$\text{I} \subset D_{\text{train}}$,	✓				
$\text{SS} \subset \mathcal{T}_{\text{train}} \backslash D_{\text{train}}$,	✓	✓		✓	
$\text{LS} \subset \{(x, y) : \|x\| > 10, \max(v) \leqslant 100\}$	✓	✓			✓
$\text{SL} \subset \{(x, y) : \|x\| \leqslant 10, \max(v) > 100\}$	✓		✓	✓	
$\text{LL} \subset \{(x, y) : \|x\| > 10, \max(v) > 100\}$	✓		✓		✓

Examples from HINT

Train		2X5÷9 2	(9-9)X(3-4)-1X(0+3-(6-(9-2÷2))) 0
		5X5+(9-0-2) 32	4X(3+9)-7-(0-5) 41
Test	I	1÷4 1	1X(2÷5)X(8-8-6) 0
	SS	1+3÷4 2	3X(7X2)+(8+4)+4X3 66
	LS	3X(8X(8X1)+0÷9) 192	5X(3+1X9)+(2-5)+(7X(6+5)) 135
		2X(3X(3÷6+6X(3+4X6÷(1X6)))+0÷3) 438	
	SL	(6X5-0)÷((4+3+5)÷9)+(3-((2-(2+(5X7-8÷9)))/4-9)) 18	
		6-3÷(9X(9÷(4-7)))+(1+1/(5-1))-(7X2+6÷8) 6	
		(7+3)/(6-6X(0X(6÷1)))-(3X1-6-4/(4-3))X(9X3) 2	
	LL	((6+7X(1+2÷4+(1+4-0÷3)X8-(1+3X8))X((0+(2X9-0)/3)÷(8÷9)) 174	
		(3+(8+(4-7X(7+8))X(8÷4-(4-(6+5)+6)))÷(7+5X1X0)÷5 1	
		7X(8÷(1X(7-4))+(1+2))X(0+9-5÷(8+4÷(9X6)))+(8-(9-8+3)) 620	

Experimental Results

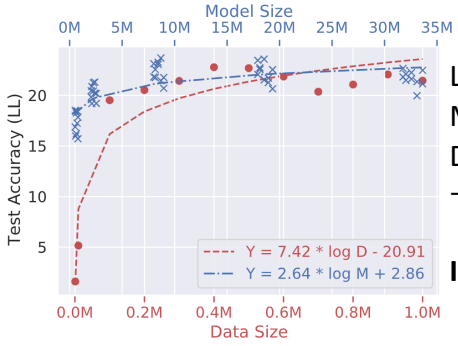
Image Inputs

Model	Variant	I	SS	LS	SL	LL	Avg.
GRU	w/o att	61.3±1.4	53.3±1.7	30.5±1.2	9.2±0.2	11.9±0.5	33.2±0.9
	w/ att	66.7±2.0	58.7±2.2	33.1±2.7	9.4±0.3	12.8±1.0	35.9±1.6
LSTM	w/o att	80.0±5.7	76.2±7.4	55.7±8.2	10.9±0.6	19.8±2.6	48.6±4.9
	w/ att	83.9±0.9	79.7±0.8	62.0±2.5	**11.2±0.1**	**21.0±0.8**	51.5±1.0
Transformer	vanilla	20.9±0.4	9.3±0.2	5.7±0.3	1.5±0.3	2.9±0.5	8.3±0.3
	rel.	86.2±0.9	83.1±1.3	60.1±2.3	10.9±0.2	19.4±0.5	51.7±1.0
	rel. uni.	**88.4±1.3**	**86.0±1.3**	**62.5±4.1**	10.9±0.2	19.0±1.0	**53.1±1.6**

Symbol Inputs

Model	Variant	I	SS	LS	SL	LL	Avg.
GRU	w/o att	74.9±1.6	68.1±0.5	42.1±1.9	10.5±0.2	14.0±0.8	41.3±0.6
	w/ att	76.2±0.6	69.5±0.6	42.8±1.5	10.5±0.2	15.1±1.2	42.5±0.7
LSTM	w/o att	84.3±5.2	79.6±6.0	63.7±6.1	11.7±0.3	22.1±1.4	52.3±3.8
	w/ att	92.9±1.4	90.9±1.1	74.9±1.5	**12.1±0.2**	**24.3±0.3**	58.9±0.7
Transformer	vanilla	93.9±0.3	91.0±0.5	33.2±1.2	11.5±0.1	11.5±0.7	47.4±0.4
	rel.	96.6±0.3	95.1±0.4	72.1±1.5	11.8±0.2	22.3±0.6	59.4±0.5
	rel. uni.	**98.0±0.3**	**96.8±0.6**	**78.2±2.9**	11.7±0.3	22.4±1.1	**61.5±0.9**
GPT-3	0-shot	19.0	9.0	3.0	10.0	2.0	8.6
	0-CoT	42.0	36.0	5.0	**49.0**	6.0	27.6

Scaling laws w.r.t. model and dataset



Log-linear scaling trend
Model: $\mathbf{10^{33}}$ parameters
Dataset: $\mathbf{10^{15}}$ examples
→ Accuracy: **90%**.

Impossible!

Few-shot Learning and Generalization

Concept	Perception	Syntax	Semantics
x		Number	11
y		Number	12
a		Op1	$\max(i, j)$
b		Op1	$\min(i, j)$
c		Op2	$(i + j) \div 2$
d		Op2	$2i \times j \div (i + j)$