Generalizing Reasoning Problems to Longer Lengths

Changnan Xiao [1] and **Bing Liu** [2]

[1] ChangnXX.github.io, changnanxiao@gmail.com

[2] University of Illinois Chicago, liub@uic.edu

Length generalization problem in learning to reason

- **Length Generalization (LG)** (or **length extrapolation):**
  - when trained on reasoning problems of smaller lengths/sizes, e.g., 345 + 67,
  - the model struggles with problems of longer lengths, e.g., 1234 + 56789.
- A popular solution to improve reasoning is to use Chain of Thought (CoT) (Wei et al., 2022),
  - CoT: providing intermediate reasoning steps
  - However, Dziri et al. (2023) and others have shown that even with detailed CoT steps, the learned models still fail to generalize.

Three main contributions

1. Present a **theorem** to identify the **root cause of LG.**
   - It explains why existing approaches are insufficient.
2. Propose a **sufficient condition for LG**, **(*n*, *r*)-consistency**
3. **Validate the theory** by learning math reasoning tasks like *arithmetic*, *parity*, *addition*, *multiplication*, and *division* to achieve LG.

Root cause of LG

- We use problem length as the dimension of a function
- **Theorem**: for a function g_N of a lower dimension *N*, there exist infinitely many continuations $f_{N'}$ of a higher dimension *N′* (*N’*> *N*) that can achieve the effect of g_N.

Theorem 3.1 *Define V as a metric space. Denote* $0 \in V$ *to be the empty token. For* $g_N : V^N \to [-1,1]$, $\forall N' > N$, *there exists infinitely many continuations* $f_{N'} : V^{N'} \to [-1,1]$ *s.t.* $f_{N'}(v_1, \ldots, v_N, 0, \ldots, 0) = g_N(v_1, \ldots, v_N)$.

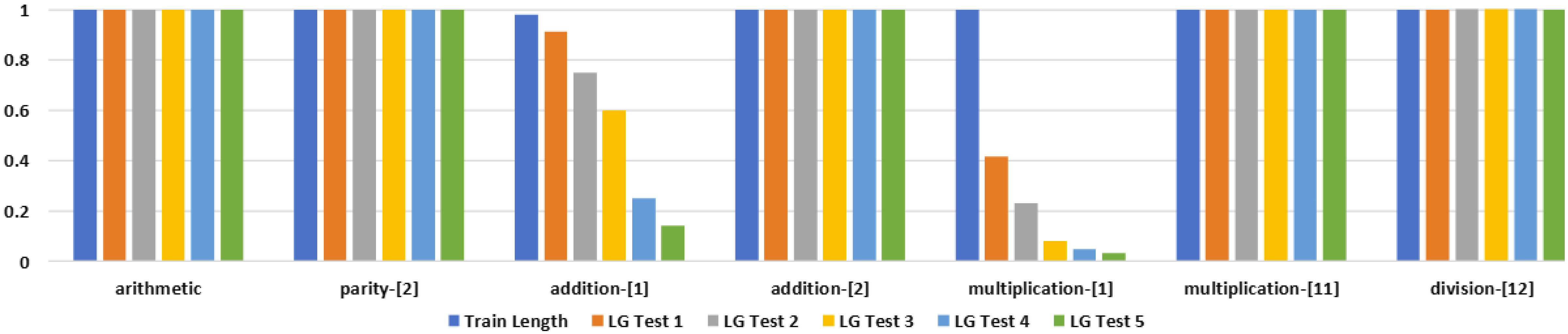
- **Implication**: Sufficient bias needs to be introduced to the problem *s.t.* with the bias, $f_{N'}$ is made equal to $g_{N'}$ **uniquely**.

Sufficient condition for LG: *(n, r)-consistency*

- **(*n*, *r*)** defines a context in a CoT step with *n* *r*-length intervals (subsequences) in the sequence (input or output expression)
- **Consistency**:
  - **(1)** the context is independent of the distances between any pair of intervals
  - **(2)** the same/consistent output should be predicted if the input of a CoT step of any instance of the problem contains the context.
- **Prove**: if a reasoning problem's CoT scheme is (*n*, *r*)-consistent, LG can be achieved with a Transformer.

Experiment settings and results

	Train Length	LG Test 1	LG Test 2	LG Test 3	LG Test 4	LG Test 5
arithmetic in F_7	$L \in [3, 20)$	$L \in [3, 30)$	$L \in [3, 40)$	$L \in [3, 50)$	$L \in [3, 60)$	$L \in [3, 100)$
parity-[2]	$L \in [1, 8)$	$L \in [1, 30)$	$L \in [1, 40)$	$L \in [1, 50)$	$L \in [1, 60)$	$L \in [1, 100)$
addition-[1/2]	$L \in [1, 8)$	$L \in [1, 9)$	$L \in [1, 10)$	$L \in [1, 11)$	$L \in [1, 16)$	$L \in [1, 21)$
multiplication-[1/11]	$L \in [1, 6)$	$L \in [1, 7)$	$L \in [1, 8)$	$L \in [1, 9)$	$L \in [1, 10)$	$L \in [1, 11)$
division-[12]	$L \in [1, 6)$	$L \in [1, 7)$	$L \in [1, 8)$	$L \in [1, 9)$	$L \in [1, 10)$	$L \in [1, 11)$



Thank You

If you are interested in the topic, please read our paper.

Three main contributions

1. Present a **theorem** to identify the **root cause of LG.**
   - It explains why existing approaches are insufficient.
2. Propose a **sufficient condition**, **(*n*, *r*)-consistency**, for LG,
   - define a problem class whose problems can have CoT schemes satisfying the (*n*, *r*)-consistency condition.
   - prove that LG can be achieved with a Transformer for this class of problems.
3. **Validate the theory** by learning math reasoning tasks like *arithmetic*, *parity*, *addition*, *multiplication*, and *division* to achieve LG.

(n, r)-consistency using **addition** as an example

- **Addition** using a 1-dimensional CoT scheme, called *addition*-[1].
- **CoT1**: input S^0 = '123+567=_**$**0' and output S^1 = '123+567=?**9**0',
  - where ? indicates 0 is carried and $ indicates 1 is carried.
  - An example (3, 3) context ('_**$**0': '123', '567') in S^0.
    - We want to predict the value for the position of $ in S^1, i.e., **9**.
- **CoT2**: S^0 = '12342+45678=_**$**0' and S^1 = '12342+ 45678=$**2**0'
  - ('_**$**0': '123', '567') also exists in this CoT
- Not (3, 3)-consistent because for the same context, the predictions are different, CoT1 is **9** but CoT2 is **2**.

An (n, r)-consistency example (cont.)

- We now use a 2-dimensional CoT scheme, *addition*-[2]
  - Adding tags as the second dimension, indicating the positions involved in the next computation step
    - CoT1: $$123 + 567 = \ \$0 \Rightarrow \begin{pmatrix} 123 + 567 = & \$0 \\ I \quad\quad J & K \end{pmatrix}, \ and$$
    - CoT2: $$12342 + 45678 = \ \$0 \Rightarrow \begin{pmatrix} 12342 + 45678 = & \$0 \\ I \quad\quad\quad J & K \end{pmatrix},$$
  - This CoT scheme is (3, 3)-consistent because the same context in any problem instance will give the same prediction for the intended position, e.g., \$. $\begin{pmatrix} \text{`}\$0\text{'} : & \text{`}123\text{'}, & \text{`}567\text{'} \\ K & I & J \end{pmatrix}$