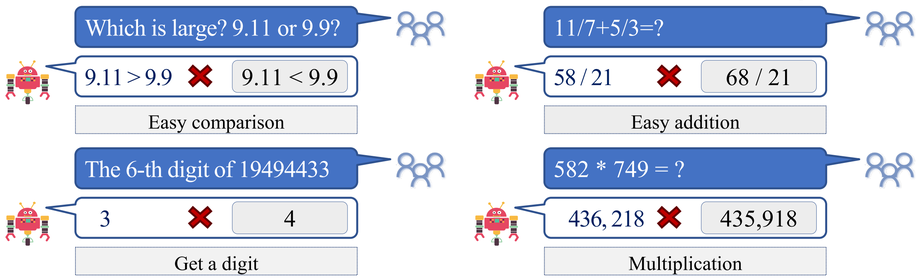
Number Cookbook: Number Understanding of Language Models and How to Improve It

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, Muhan Zhang

paper github

PEKING UNIVERSITY ICLR

Motivation: LLMs cannot handle numbers perfectly.



Real responses from GPT-4o

Current benchmarks do not focus on numbers.

Current benchmarks cannot test the practical **N**umber **U**nderstanding and **P**rocess **A**bility (**NUPA**) of LLMs.

- Mixed up with math reasoning. (Numbers are not equal to math!)
- The numbers are simplified and impractical (like short integers, very short fraction and so on).

GSM8k	Question: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? Answer: Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May. #### 72
MATH	Question: If $\sum_{n=0}^{\infty}\cos^{2n}\theta = 5$, what is $\cos 2\theta$? Answer: The geometric series is $1+\cos^2\theta+\cos^4\theta+\cdots=\frac{1}{1-\cos^2\theta}=5$. Hence, $\cos^2\theta=\frac{4}{5}$. Then $\cos 2\theta = 2\cos^2\theta - 1 = \frac{3}{5}$.

NUPA benchmark focuses on **number itself**.

- **Practical**: selected from primary and middle school textbooks.
- **Comprehensive**: 4 kinds of number representations and 17 tasks.

4 number representations

- Integer: 123, 12, …
- Float: 3.1415926 …
- Fraction: 7/11, 4/3, …
- Scientific Notation: 3.1e2, …

17 tasks in 4 categories

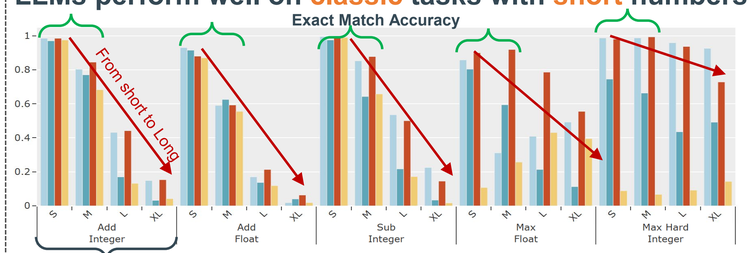
- Elementary arithmetic: +,−,×,/,//,%
- Compare: >, <
- Digit understand: `len`, `get_digit`, `count`,…
- Conversion: Float2Scient, Scient2Float, Sig. Fig., …

4 × 17 (with some filtering)= Overall 41 tasks

	Elementary Arithmetic						Compare		Digit Understanding						Conversion		
	Add	Sub	Multi-ply	True div	Floor div	Mod	Max	Min	Digit Max	Digit Min	Digit Add	Get Digit	Length	Count	To Float	To Scient	Sig. Fig.
Integer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Float	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓			✓	✓
Fraction	✓	✓	✓	✓			✓	✓							✓		
Scientific	✓	✓	✓				✓	✓							✓		

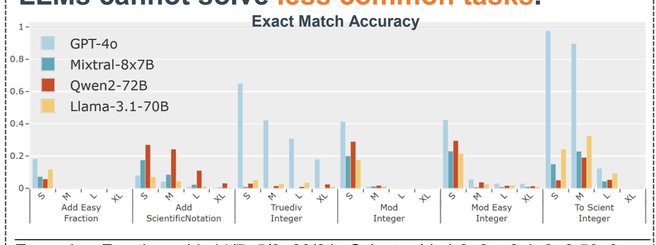
Current LLMs show restricted number ability.

LLMs perform well on classic tasks with short numbers



but **length** is still an obstacle to number processing.

LLMs cannot solve less common tasks.



Examples: Fraction add: 11/7+5/3=68/21 Scient add: 1.3e2 + 2.4e3=2.53e3
Int truediv: 12/9=4/3 Int mod: 142%58=4
Int mod(easy): 142%7=2 Int to_scient: 15325=1.5325e4

Digit-related tasks are easy for humans but difficult for LLMs.

Examples: get_digit(31415926, 5)=5 To scient: 12874000=1.2874e7
length(31415926)=8 Count: count(12867192**7**,7)=2
digit_max(837**1**2,184**95**)=88795

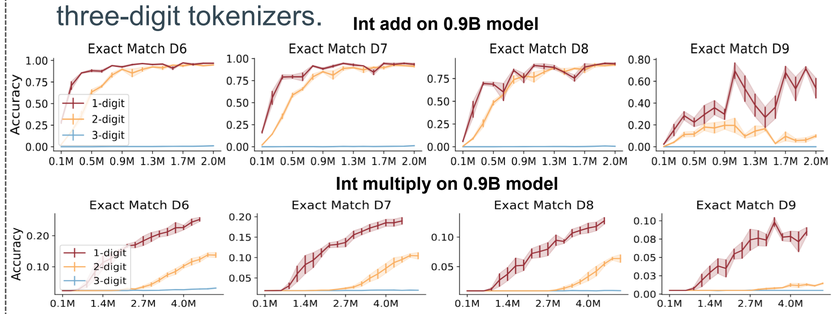
- **Easy for humans even with a long number.**
- **LLMs cannot accurately answer the questions, especially for longer numbers.**

How to improve NUPA?

Number related techniques?

Train from scratch with different *tokenizers*, *positional encoding (PE)*, and *data formats* on different NUPA tasks. Train on 1-8 digits (called *in-domain*), test on 1-20 digits (9-20: *out-of-domain*).

- **Tokenizers**: the one-digit tokenizer is better than two- or three-digit tokenizers.

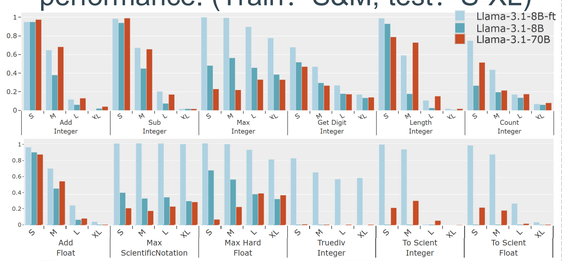


X-axis: seen training samples, y-axis: exact match accuracy; Dm means the longer input has m digits. More experiments in paper.

- **PE:** Weaker or no PEs slow down in-domain learning but bring out-of-domain generalization for all tasks and model sizes.
- **Data formats**: integer-part reversed representation (like 123.45 → 321.45) slightly helps digit alignment.

Post-training?

- Direct finetuning on NUPA tasks improve performance. (Train: S&M, test: S-XL)



- Techniques (tokenizers, PE, data formats) cannot be directly used in post-training.

	Integer Addition				Float Addition			
	S	M	L	XL	S	M	L	XL
Direct finetuning	**0.95**	**0.65**	**0.12**	0.01	**0.96**	**0.71**	**0.27**	**0.08**
w/o Finetuning	**0.95**	0.38	0.06	**0.02**	0.90	0.47	0.10	0.02
NoPE	0.67	0.04	0.00	0.00	0.37	0.06	0.00	0.00
NoPE + rev + 1d	0.89	0.35	0.06	**0.02**	0.81	0.38	0.09	0.01
NoPE + rev + pad + 1d	0.87	0.34	0.05	**0.02**	0.74	0.38	0.06	0.01
RoPE + 1d	0.93	0.59	0.05	0.00	0.33	0.30	0.06	0.01
RoPE + rev + 1d	0.40	0.20	0.04	0.00	0.35	0.30	0.09	0.02

Change PE data format tokenizer

Chain-of-thought?

- Good performance but has much longer context and lower response speed (17×).

From now on, there is no general solution for number tasks!